Chromatin states and architecture during stem cell differentiation: analyses and predictions

Akshay Shah

Thesis for the degree of Philosophiae Doctor (PhD)

Department of Molecular Medicine
Institute of Basic Medical Sciences
Faculty of Medicine
University of Oslo, Norway

2015

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo
No. 2129

ISBN 978-82-8333-157-8

Cover: Hanne Baadsgaard Utigard.
Print production: John Grieg AS, Bergen.

Produced in co-operation with Akademika Publishing.
The thesis is produced by Akademika Publishing merely in connection with the thesis defence. Kindly direct all inquiries regarding the thesis to the copyright holder or the unit which grants the doctorate.

Contents

Acknowledgements

The work included in this thesis was funded by a PhD stipend from the Molecular Life Science program of the University of Oslo. It was supervised by Professor Philippe Collas at the Department of Molecular Medicine, University of Oslo.

Most importantly, I would like to thank my supervisor, Philippe Collas. I will always be grateful for the opportunity you provided and the knowledge you imprinted. It has been inspiring to see you still work harder than a PhD student and it will remain a motivation for me to follow in your footsteps. Your enthusiasm and unmatchable passion for research has been the force that has pushed me to outperform myself. I thank you for all the biology lessons, patient explanations and for upgrading my popular science-like writing to high-level scientific literature.

I would like to thank Jan Øivind and the whole of Collas lab, especially for not complaining about my loud laughter. In fact, I thank them to be a reason for the happy memories. Andrew, Eivind, Jonas and Monika helped tremendously with all the bioinformatics analyses and on numerous occasions, lit the light bulb required to proceed with the research. Anita, Anja, Erwan, Kristin and Thomas patiently guided me with lab experiments even though I had to end up hanging the lab-coat for working on the computer. Sumithra, the Asian connection, was ready to tackle the administration and impart wisdom on the comings and goings in Oslo.

I am thankful to my fellow PhD cronies, Graciela, Jane, Kristina and Torunn for all the lunch time planning debacles and the regular weekend beers and, of course, the insightful discussions on merging our research topics.

I owe a deep sense of gratitude to my family, my aunt and uncle, and my soon-to-be wife, Tara. My mother restrained her Bollywood-esque feelings to let me continue working away from home and my father encouraged me throughout the way while my sister made sure I did not make many mistakes along the way. My aunt (Sushilaben) and uncle (Virchand) guided and helped during my time in the UK. Tara took away my true love, cookies and fried food, to make sure I remained healthy while working towards my thesis and provided unconditional encouragement and support.

List of Publications

Paper I
A. Shah, A.R. Oldenburg and P. Collas. A hyper-dynamic nature of bivalent promoter states underlies coordinated developmental gene expression modules. *BMC Genomics*, 15(1):1186, 2014

Paper II
T. Rønningen, A. Shah, A. Oldenburg, K. Vekterud, E. Delbarre, J. Moskaug and P. Collas. Pre-patterning of differentiation-driven nuclear lamin A/C-associated chromatin domains by GlcNAcylated histone H2B. *Genome Research*, 2015. *Accepted*.

Paper III
A. Shah, J. Paulsen and P. Collas. Inference of promoter-enhancer contacts from epigenomics datasets reveals dynamic interaction during adipogenic differentiation. *Manuscript*

List of Abbreviations

3C	Chromatin conformation capture
4C	Circular chromosome conformation capture
5C	Chromosome conformation capture carbon copy
ac	Acetylation
ASC	Adipose-derived stromal/stem cells
BIC	Bayesian information criterion
bp	base pair
BWT	Burrows-Wheeler Transform
ChIA-PET	Chromatin interaction analysis with paired-end tag sequencing
ChIP	Chromatin immunoprecipitation
DBN	Dynamic Bayesian Network
DNA	Deoxyribonucleic acid
EDD	Enriched domain detector
EST	Expressed sequence tag
FDR	False discovery rate
GAD	H2BS112GlcNAc domain
GlcNAc	O-GlcNAcylation
HBP	Hexosamine biosynthetic pathway
HMM	Hidden Markov Model
HTS	High-throughput sequencing
K	Lysine
kb	Kilobase
LAD	Lamina-associated domain
Mb	Megabase
me	Methylation
miRNA	Micro ribonulceic acid
ncRNA	Non-coding ribonulceic acid
OGA	O-GlcNAcase
OGT	O-GlcNAc transferase
OOB	Out-of-Bag
P-E	Promoter-enhancer
PCR	Polymerase Chain Reaction

PolyA	Polyadenylation
PTM	Post-translational modification
RNA	Ribonulceic acid
RNAPII	RNA polymerase II
RF	Random Forest
SNP	Single nucleotide polymorphisms
TAD	Topologically-associated domain
TF	Transcription factor
TSS	Transcription start site
S	Serine
seq	Sequencing

Chapter 1

Introduction

Genetic information is encoded in deoxyribonucleic acid (DNA), a double-stranded polymer of four bases (adenine (A), thymine (T), guanine (G) and cytosine (C)). The two strands are paired by hydrogen bonding between A-T and G-C base pairs (bp) [1]. The human genome consists of approximately three billion bp within 23 chromosomes altogether containing $\sim$2 m of linear DNA. In order to fit into a $\sim$10 µm diameter nucleus, DNA must be folded and packaged into a highly organized structure, the fundamental unit of which is the nucleosome. Nucleosomes form the basic structure of chromatin. Histones H2A, H2B, H3 and H4 serve as a scaffold and regulatory proteins; winding up DNA into arrays of particles called nucleosomes [2]. Histones are positively charged and bind tightly to the negatively charged DNA [3]. Each nucleosome wraps $\sim$1.7 turns, or $\sim$146 bp, of DNA in an array of repeats [4] that appear as 'beads on a string' by electron microscopy [5]. Nucleosomes are spaced by the linker histone H1 or its variant histone H5 [2, 4].

Approximately 2% of the DNA consists of protein-coding genes [6]. However, the intergenic regions encode functional non-coding ribonucleic acid (ncRNA) [7] and consist of other regulatory elements such as enhancers [8]. Enhancers can also be found in the intragenic regions of genes [9]. Enhancers contribute to regulating the strength of gene expression by interacting with promoters, which are spatially restricted regions around the transcription start site (TSS) of genes [8] (Figure 1). As we will address later in this thesis, promoters and enhancers harbor many combinations of epigenetic modifications which affect their activity, their interaction

with transcriptional regulators and their localization in the nucleus. The end-result of these modifications is a tightly regulated spatial and temporal regulation of gene expression, driving developmental and cell differentiation programs and cell and tissue homeostasis.

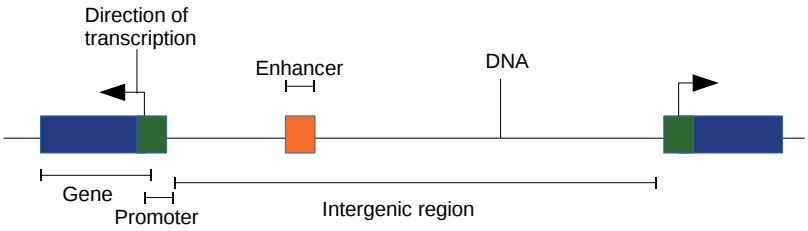

Figure 1: Schematic annotation of the genome into genes, promoters, enhancers and intergenic regions.

1.1 Epigenetic regulation of chromatin function and conformation

Epigenetics refers to the study of processes taking place "on top of" (epi- in Greek) the DNA code: these processes consist of heritable biochemical modifications of DNA or chromatin which affect gene expression but not the underlying DNA sequence [10]. Epigenetics represents a "link" between genotype and phenotype. The outcome of a particular gene (repressed or expressed at various levels) is represented in Waddington's epigenetic landscape proposed in 1957 (Figure 2). The notion is that a cell makes numerous decisions while differentiating and these decisions lead to a particular cell fate producing a particular cell type. In the last two decades, many factors contributing to epigenetic regulation of gene expression have emerged and are increasingly being understood.

1.1.1 Post-translational modifications of histones

Histones are subject to numerous post-translational modifications (PTMs). PTMs are found in the globular domain and in the flexible N-terminal tail of all his-

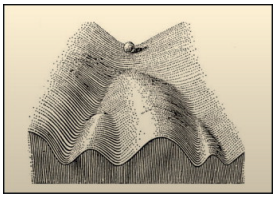

Figure 2: Waddington's Classical Epigenetic Landscape. This is the metaphorical visualization of a cell (represented by the ball) capable of taking different permitted trajectories leading to specific outcomes in cell fate. Figure taken from [11].

tones. PTMs include acetylation and methylation of lysines (K) [12] and arginines (R) [13, 14], phosphorylation of serines (S) [15] and threonines (T) [16], ubiquitylation and sumoylation of lysines [17], and β-N-acetylglucosamination of serines and threonines [18] (Figure 3). Additional PTMs include ADP ribosylation, ubiquitination, deamination, proline isomerisation, biotinylation and cleavage of the histone tails [19–21].

Acetylation on lysine residues is generally coupled to a transcriptionally permissive state, probably by neutralizing the basic charge of lysines in the N-terminal tail of histone H3 and H4, relaxing DNA-histone interactions or histone-histone interactions. In contrast, mono- and di-methylation of arginine and mono-, di-, and tri-methylation of lysine can repress or activate transcription depending on the residue modified [22]. Although considered as a mark of active genes, H3 lysine 4 trimethylation (H3K4me3) occupies the promoter and TSS of many genes irrespective of their transcriptional status [23] and thus should be regarded as a promoter mark. Nonetheless, H3K4me3 has also been reported on gene bodies in Arabidopsis [24], zebrafish [24] and humans [25] in association with transcriptional activity. H3K4me2 is more typically associated with transcriptionally active sites in chromatin and shows a dynamic distribution on promoters, introns and intergenic regions [26]. H3K4me1 has been found to mark enhancers regardless of the expression status of their cognate genes [27]. Co-enrichment of H3K4me1 with acetylated H3K27 (H3K27ac) marks enhancers of active genes [27]. In contrast, in mammalian cells H3K27me3 is enriched on inactive promoters [28] but can also be distributed broadly across inactive regions of the genome [29].

Many active histone PTMs associate with transcriptional complexes and RNA

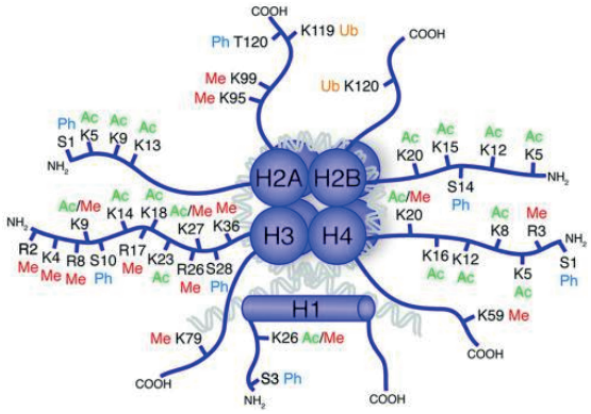

Figure 3: Nucleosomal histones and post-translational modifications on their N-terminal tails, including methylation (Me), acetylation (Ac), ubiquitination (Ub), and phosphorylation (Ph). The N-terminal tails of histone tend to protrude from the nucleosome core, making them accessible to "writers" and "readers" of the epigenetic code. Adapted from [31].

polymerase II (RNAPII). While the H3K4 histone methyltransferase (HMT) SET1 is targeted to the 5' end of genes through recruitment by the initiating serine (Ser)5-phosphorylated RNAPII [30], the H3K36 HMT SETD2 associates with the Ser2-phosphorylated elongating form of RNAPII. Hence, H3K4me3 marks the TSS and 5' end of genes, while H3K36me3 marks transcribing gene bodies and the 3' end of genes [29].

1.1.2 Combinations of histone modifications form "chromatin states" that modulate transcription

Histone modifications modulate transcription outcome by regulating the chromatin conformation or by blocking or recruiting transcriptional regulators. Histone PTMs can lead to varying degrees of chromatin compaction. The relatively "open" chromatin conformation is referred to as euchromatin – which appears as electron-light by electron microscopy, whereas the compact form of chromatin is referred to as heterochromatin – appearing as electron-dense in the electron microscope (Figure 4). Euchromatin is overall gene-rich (i.e. its gene density is

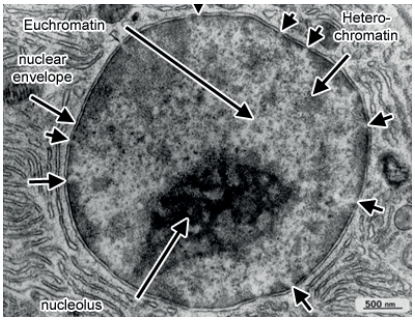

Figure 4: Electron micrograph of oligodendroglial satellites of neuron in the prefrontal cortex of the brain. Heterochromatin appears more electron-dense (darker) than euchromatin and tends to be located near the nuclear periphery. Abbreviations: N, nucleus; H, heterochromatin; E, euchromatin; Neu, neuron; M, mitochondria; R, reticulum; Rib, ribosomes. Taken from [45]

higher than that of the genome average gene density of $\sim$8 genes/megabase (Mb) in humans). Heterochromatin in contrast is gene-poor and mostly contains transcriptionally inactive protein-coding genes [32–34]. Increasing evidence indicates however that many genes encoding ncRNAs are transcribed from heterochromatic regions [35–37]. Heterochromatin is predominantly found in telomeres, centromeres and pericentromeric regions [38, 39]. These domains are marked by distinct histone PTMs and are differentiated by boundary elements often referred to as insulators [40]. Heterochromatin is characterized by H3K9me3 and its effector protein heterochromatin protein 1 (HP1/CBX3) [41]. However, the finding of H3K9me3 and HP1 in coding regions of transcribed genes has challenged the view of H3K9me3 as a strict marker of heterochromatin [42]. This shows the importance of genomic context on the transcriptional outcome of histone PTMs. Another mark of heterochromatin is H4K20me3, enriched in repeat regions [38, 43, 44]. Heterochromatin is further characterized by a lack of histone acetylation, H3K4me2/3 or H3K36me2/3 [42]. In contrast, euchromatin is characterized by promoter H3K4me3, H3K36me3 on active gene bodies, H3K79me3 and various forms of lysine acetylation [22].

The advent of DNA high-throughput sequencing (HTS) technologies downstream of chromatin immunoprecipitation (ChIP), a technique referred to as ChIP-seq (see section 1.4.1) of modified histones and transcription regulators, chromatin

landscapes have started to emerge. Data from *Drosophila* reveal no less than five main "chromatin states" enriched in specific combinations of 53 chromatin binding proteins [46]. Additional studies have similarly identified a number of distinct chromatin states and have attributed these states to biological functions (e.g. active, poised, repressed domains) [47, 48]. Interestingly, combining histone PTMs with RNAPII binding data enables prediction of the activity level of regulatory regions during development [49]. As discussed later (section 1.5), bioinformatics methods have enabled biological functions to be ascribed to epigenetic modifications.

1.1.3 A relationship between cellular metabolism and chromatin organization?

Cell metabolism is coupled to histone PTMs and to the activity of chromatin remodeling proteins affecting chromatin structure and gene expression [50]. Metabolic intermediates often act as co-factors or substrates of histone modifying enzymes. Of importance for work presented in this thesis (**Paper II**), the hexosamine biosynthetic pathway (HBP) is responsive to intracellular levels of amino acids, fatty acids and carbohydrates [51] and constitutes an important link between glucose metabolism and chromatin. Approximately 3-5% of glucose taken up by the cell is directed to the HBP [52] and converted to UDP-N-acetylglucosamine (UDP-GlcNAc), the donor of O-linked GlcNAc for O-GlcNAcylation of proteins. GlcNAcylation is catalyzed by the O-GlcNAc transferase OGT [53], while O-GlcNAcase (OGA) hydrolyzes O-GlcNAc [54, 55]. All core histones can be GlcNAcylated [56–58], indicating that chromatin organization is influenced by the OGT/OGA balance.

GlcNAcylation of H2B on Ser 112 (H2BS112GlcNAc) has been reported in mammalian cell lines at the TSS of transcribed genes [18,59,60]. H2BS112GlcNAc has also been claimed to promote H2BK120 monoubiquitination (H2BK120ub1) in HeLa cells, suggesting a link to gene activity [18]. However in *Drosophila* cells, OGT has been linked to transcriptional repression [61]. In *Drosophila*, OGT is a Polycomb Group protein essential for Polycomb-mediated gene repression during development [61, 62]. OGT also modifies proteins involved in transcriptional repression [63]. For instance, GlcNAcylation of the Polycomb repressor complex

2 protein EZH2 (which methylates H3K27) regulates EZH2 protein stability, and H3K27me3 partly depends on OGT expression [64]. Moreover, GlcNAcylation of the SIN3A subunit of histone deacetylase HDAC1 has a repressive impact on gene expression [65]. Thus, it is possible that H2B-S112GlcNAc plays distinct roles in different cell types. It also emerges from these studies that the connection between gene expression and H2BS112GlcNAc is still unclear. Moreover, positioning of H2BS112 on the nucleosome surface raises the hypothesis that H2BS112GlcNAc promotes other chromatin-associated processes. This is a key premise of **Paper II**.

1.2 Towards understanding the 3D genome: associations of chromatin with nuclear lamins

The nuclear envelope contributes to defining position, shape and functions of the eukaryotic cell nucleus. The nuclear envelope consists of an outer and inner nuclear membrane underlined in the nucleoplasmic side by an intermediate filament meshwork called the nuclear lamina [66–68]. Two types of nuclear lamins make up the nuclear lamina: A-type lamins (lamins A and C, often referred to as lamin A/C) which are splice variants of the *LMNA gene*, and B-type lamins, consisting of lamins B1 and B2 encoded by the *LMNB1* and *LMNB2* genes respectively. B-type lamins are constitutively expressed and are anchored to the inner nuclear membrane through a farnesylated CAXX motif in their C-terminus [68]. In contrast, A-type lamins are developmentally regulated: they are undetectable or expressed at low levels in early embryos and embryonic stem cells but are expressed in more differentiated progenitors and terminally differentiated cells [68–70] with a few exceptions [43]. As part of their maturation, A-type lamins lose the C-terminal farnesylation site and thus do not anchor into the inner nuclear membrane [68]. Instead, they associate with B-type lamins in the peripheral nuclear lamina [71] and a pool of lamin A/C remains nucleoplasmic [72, 73].

The intranuclear fraction of lamin A/C depends on the nucleoplasmic protein lamina-associated polypeptide LAP2α which directly binds lamin A/C and chromatin [74, 75]. Nucleoplasmic lamin A/C and LAP2α affect retinoblastoma

protein function [76] and promote cell cycle arrest in tissue progenitor cells [75, 77, 78], and LAP2α overexpression in mouse pre-adipocytes favors adipogenic differentiation [79]. Nucleoplasmic lamin A/C appears therefore to be critical for differentiation of tissue progenitor cells; they may also play a role in the regulation of gene expression in the nuclear interior [73, 80]. Our laboratory and others have notably shown that lamin A/C-genome interactions can also occur on promoters [81, 82], suggesting a role of lamin A/C in gene regulation, perhaps in the nuclear interior (which is gene-rich). **Paper II** reports a large-scale spatial reorganization of chromatin during differentiation of human adipose tissue stem cells into adipocytes, involving association/dissociation events of chromatin with A-type lamins.

Presumably due to their ability to bind DNA and nucleosomes *in vitro* [83], A- and B-type lamins associate with chromatin [84]. These contacts occur through large lamina-associated domains – which for reasons apparent in **Paper II** we rename lamin-associated domains (LADs). LADs span 0.1 to 10 Mb [34, 84–91] (Figure 5). LADs are generally conserved between cell types but some cell type-specificity exists [92]. LADs are overall gene-poor and transcriptionally inactive [86, 92, 93]. Accordingly, LADs are enriched in histone modifications characterizing silent chromatin, including H3K9me2 or me3 [86, 89–91]; this is consistent with the heterochromatic environment of the nuclear periphery [43, 94–96]. Targeting and anchoring of loci at the nuclear periphery in C. elegans or mouse cells requires H3K9me2/3, H3K27me3, lamin A/C (in mammals), short DNA sequences and protein factors [96–98]. This suggests that LAD formation is tightly regulated in a temporal and cell type-specific manner.

Recent data from our laboratory and others [82, 87, 90, 99] suggest that lamin-chromatin interactions are under developmental control. Notably, in adipose tissue stem cells (ASCs – a cell type we have used in our work), pro-adipogenic gene promoters are released from lamin A/C after differentiation into adipocytes, whereas many non-adipogenic, lineage-specific promoters remain associated with lamin A/C [82]. Moreover, dissociation of a myogenic promoter from *C. elegans* lamin has also been linked to muscle-specific gene activation [95]. These results again argue that lamin-genome interactions may be under developmental regulation. This aspect is examined at the genome-wide level in **Paper II**.

Nuclear lamins play an important role in the organization of heterochromatin

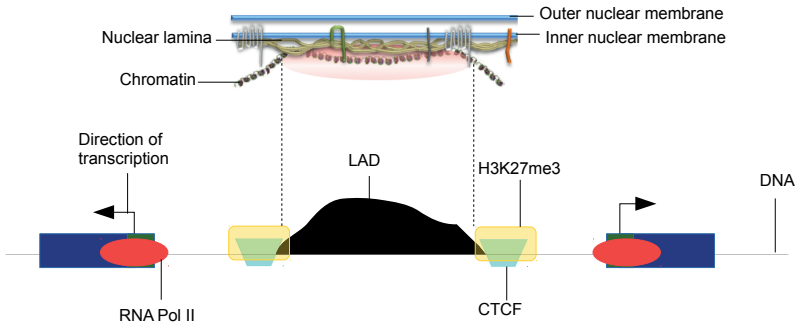

Figure 5: Schematic Representation of a lamin-associated domain (LAD). LADs are mostly gene poor and heterochromatic, and their borders are in some cases enriched in the insulator protein CCCTC-binding factor CTCF and in H3K27me3.

[43, 82, 89–91, 99, 100] and impact spatial genome conformation [34]. Lamins also anchor signaling molecules and transcription factors [101, 102] and connect the nuclear interior with elements of the cytoskeleton [68] influencing the position of cell nucleus [68, 103]. It is thus not surprising that mutations in nuclear lamins and lamin A/C in particular cause diseases. More than 400 lamin A mutations cause ∼15 diseases commonly called laminopathies, showing symptoms such as partial lipodystrophies, myodystrophies, cardio-myopathies, skeletal abnormalities or premature aging (progeria) [66,102,104]. Pathways linking lamin mutations to diseases are unknown but involve abnormalities in chromatin organization [95], signal transduction [102, 105] and autophagy [106].

A common feature of partial lipodystrophies caused by specific lamin A mutations is metabolic disorders involving glucose intolerance and insulin resistance, often leading to type II diabetes [107]. Remarkably, these syndromes are also linked to alterations of the HBP pathway and O-GlcNAC cycling (see above) due to HBP overactivity or mutations in OGA [50]. We rationalized in **Paper II** that the common metabolic disorders caused by deregulation of protein O-GlcNAcylation and lamin A-linked lipodystrophies might underline a relationship between lamin A/C and chromatin modifications modulated by H2BS112GlcNAc.

1.3 Regulation of the genome by 3-dimensional chromatin organization

1.3.1 Interactions between promoter and enhancer elements regulate gene expression

Enhancers play a critical role in regulating the strength of gene expression and the activity level of a promoter [108]. Enhancers typically consist of relatively small genomic segments (a few hundred base pairs) that harbor binding sites for protein complexes containing transcription regulators [8, 109, 110]. Enhancers exert their activity on promoters interacting with their cognate promoters through looping of chromatin [108, 111]. The enhancer nearest to a promoter in the linear genome does not necessarily regulate or interact with that promoter. Indeed, chromatin conformation capture (3C) studies and other techniques derived from 3C, such as chromatin interaction analysis with paired-end tag sequencing (ChIA-PET), reveal that promoter-enhancer (P-E) interactions often occur over tens of kilobases up to megabases apart in the linear genome [112]. Enhancers may even loop over nearby promoters (and not interact with them) to make physical contact with a more distant cognate promoter [112, 113]. In addition, an enhancer influencing a given gene can either be upstream or downstream of the gene. Therefore, identifying which enhancer regulates activity of which promoter does not merely imply searching for the nearest enhancer. Identifying P-E interactions in a 3-dimensional (3D) context is essential.

Several studies have attempted to identify criteria for interaction of an enhancer with a promoter, which have led to a general consensus of mechanisms that may be responsible for mutual selectivity of interacting promoters and enhancers [113–116] (Figure 6). These mechanisms include biochemical compatibility, spatial architecture, insulation and chromatin environment. Biochemical compatibility (Figure 6A) refers to a condition where a given enhancer "E1" is compatible with a promoter "P" while another enhancer "E2" is incompatible. Spatial architecture (Figure 6B) refers to a situation where both E1 and E2 enhancers are compatible with promoter P, but only E2 can interact due to the spatial architecture it is in [117, 118]; for instance, a promoter may be in a different topologically-associated domain (TAD; see also section 1.3.2) from an enhancer

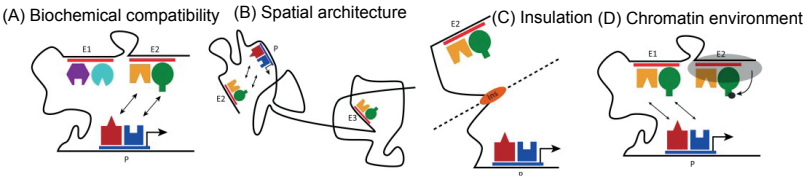

Figure 6: Distinct mechanisms can drive promoter–enhancer interaction specificity. (A) Biochemical compatibility. (B) Spatial architecture. (C) Insulation. (D) Chromatin environment. See text for details. Adapted from [8]).

it could potentially interact with, but it cannot interact because of physical constraints [116, 119]. However, interaction might, for instance, become possible following a mutation causing a conformational switch. Recent evidence of single nucleotide polymorphisms (SNPs) causing mutations also eliciting a switch in P-E interaction support the spatial architecture model of P-E interaction [120, 121]. In the case of insulation (Figure 6C) as a factor affecting a P-E interaction, enhancer E2 cannot interact with promoter P even though P and E2 are compatible, due to the presence of an insulator element between the two (e.g. CTCF; see below). Chromatin environment (Figure 6D) refers to epigenetic modifications at enhancers, promoters and intervening sequences which may modulate P-E interactions. Importantly, the chromatin context of enhancer E2 may also influence whether promoter P and enhancer E1 can interact [122, 123].

Advances in HTS techniques (section 1.4) greatly facilitate the study of genome and chromatin architecture at the genome-scale level [124–126]. However, generating genome-wide data for many cell types with high resolution remains exhaustive and expensive. Therefore, there is strong motivation to develop methods to predict P-E interactions using existing information from chromosome interaction and epigenomic datasets. In **Paper III**, we report an epigenetic pipeline to identify P-E interactions in cell types for which chromosomal interaction data do not exist.

1.3.2 CTCF

The CCCTC-binding factor CTCF [127] has been described as an insulator element shown to be a key regulator of P-E interactions [113, 128]. It contains 11

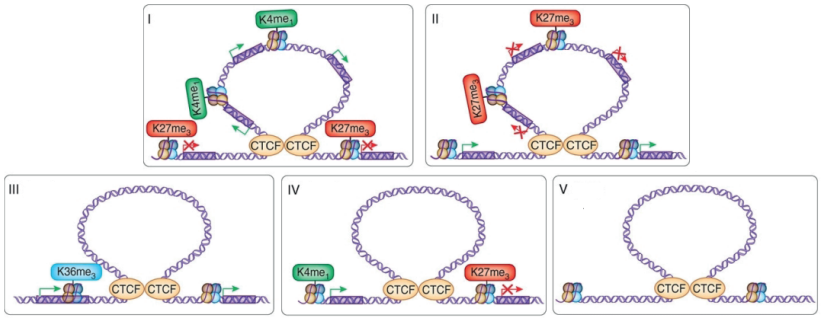

Figure 7: Proposed functions of CTCF. (I) CTCF-mediated chromatin loops contain active genes (green boxes) and H3K4me1 in the loop region and inactive genes (red boxes) and repressive marks such as H3K27me3 outside the loop. (II) Inactive genes and repressive marks inside the loop and active genes outside the loop. (III) H3K36me3 outside the loop on one side. (IV) Active genes and H3K4me1 outside the loop on one side and H3K27me3 outside the loop on the other side. (V) Chromatin loops do not appear to contain any characteristic gene expression or histone modification patterns. Figure adapted from [135]

zinc-fingers that constitute highly conserved DNA-binding domains [129]. CTCF binds the mammalian genome at 55,000-65,000 sites [130]; of which ∼23,000 binding sites are constitutive in 123 ENCODE cell types [131], whereas 30-60% are cell-type specific [132].

CTCF emerges as a key player in the spatial organization of the genome by organizing long-range chromosomal interactions (Figure 7). Circular chromosome conformation capture (4C) studies indicate CTCF binding sites are required on the H19 imprinting control region (ICR) to form inter- and intra-chromosomal interactions [133]. CTCF binding at DNase I hypersensitivity sites is also required to maintain chromatin architecture at the *Hbb* locus in mice [134]. CTCF is important in organizing spatial genome architecture (Figure 7). CTCF tethers DNA strands on separate loci, thus forming CTCF-mediated chromatin loops. CTCF is involved in the formation of chromatin loops to isolate active chromatin regions from inactive ones and isolate regions with distinct chromatin states. These features of CTCF constituted the premises for using CTCF binding data in our study of P-E interactions (**Paper III**).

A role of CTCF as an insulator protein has emerged from studies suggesting that CTCF acts as a barrier to prevent spreading of heterochromatin 'seeded' at

silenced integrated transgenes [127]. CTCF however does not systematically insulate repressed chromatin domains: only 2-4% of H3K27me3 domain borders contain CTCF in HeLa cells [132], and similarly, ∼9% of LAD borders in fibroblasts harbor CTCF [86]. CTCF also appears to be responsible for creating functional gene expression domains in which CTCF loops contain marks of active chromatin while repressive modifications are kept outside the loop [136]. CTCF may also be involved in the establishment of chromatin domains where gene expression is regulated in cohorts [137].

CTCF has also been characterized as an enhancer blocker [128, 138–141]. However, this seems to be at specific loci only [128, 138, 139]. CTCF may also positively regulate P-E interactions [142–145]. Using chromosome conformation capture carbon copy (5C), it has been shown that 79% of long range interactions are not blocked by CTCF. However, a portion of the distal interactions are enriched in CTCF and/or enhancer marks [113]. CTCF-mediated interactions are established between enhancers and their cognate promoters prior to transcriptional upregulation [146, 147] and ChIP-seq studies point to CTCF-mediated targeting of cognate promoters by regulatory elements enriched in CTCF-binding sites [148].

Lastly, CTCF has been linked to the formation of topologically-associated domain (TAD) borders [149–152]. However, with one study showing that only 15% of CTCF-binding sites are at TAD borders and 85% within the TADs [119]. Thus, CTCF within TADs might be responsible for blocking or facilitating intra-TAD P-E interactions.

1.4 Sequencing Techniques to study genome-wide chromatin regulation

1.4.1 ChIP-seq

A plethora of proteins interact with DNA to regulate functions such as transcription, DNA replication, DNA repair, chromosome organization. ChIP is to date, arguably, the method of choice to identify genomic locations of these protein-DNA interactions [153]. ChIP has been widely used to map the location of histone PTMs, transcription factors (TFs) and other non-histone proteins which interact

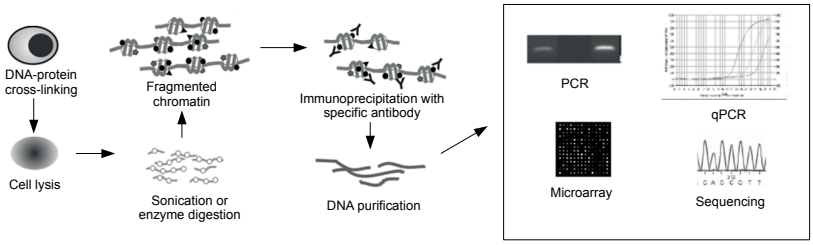

Figure 8: An outline of the ChIP protocol. The steps include DNA-protein cross-linking, cell lysis, shearing of chromatin, immunoprecipitation with a specific antibody, DNA purification and mapping of precipitated DNA by various methods. Figure adapted from [157]

with chromatin (in)directly. Identification of chromatin/DNA protein binding sites can be confined to specific genomic sites or extended to a genome-scale.

ChIP (Figure 8) consists in reversibly cross-linking DNA and proteins, most commonly using formaldehyde. This maintains association between DNA and proteins throughout the chromatin preparation and immunoprecipitation steps. However, it is not necessary to cross-link DNA and proteins to investigate histone PTMs (this is in the case called "native ChIP") as DNA binds to histones tightly [154, 155]. Chromatin is sheared to $\sim$200-500 bp fragments, insoluble complexes are sedimented and resulting chromatin is used for immunoprecipitation with an antibody against the protein of interest. The ChIPed material is washed stringently, cross-linking is reversed by high temperature (68^{o}C), proteins are digested and ChIP DNA isolated. This DNA contains genomic segments in contact with the protein of interest, which can be identified by polymerase chain reaction (ChIP-PCR), hybridization to microarrays (ChIP-chip) or by HTS (ChIP-seq) [156].

Various histone PTMs display distinct enrichment profiles throughout the genome [28, 29, 158, 159]. Transcription factors typically display a sharp binding pattern over a restricted genomic site and some histone PTMs (e.g. H3K4me3) tend to display a narrow binding pattern. Thus, genomic enrichment of such proteins can be identified without a sequenced control (reference) chromatin sample [160]. Typically however, a reference sample is required for proteins with wider binding patterns because their enrichment level over a genome-average may not always be prominent. Control samples for ChIP analyses can be cross-linked and fragmen-

ted chromatin that is not subject to immunoprecipitation ("input" chromatin) or ChIP using a non-specific antibody.

1.4.2 High-throughput sequencing-based techniques to study chromatin organization in 3D

ChIP provides a snapshot of proteins bound to chromatin. However, alongside, the 3D architecture of the genome plays an important role. 3C was invented by Dekker and colleagues [161]. This method originally relied on quantification by PCR to assess interaction between two pre-determined genomic sites, but has increasingly begun to be replaced by sequencing to provide a genome-wide view of chromosomal interactions.

3C is a method to isolate DNA fragments in spatial proximity and in-turn analyze interaction between two loci. Figure 9 illustrates the basics of 3C-based technologies. The steps in 3C include cross-linking of DNA with formaldehyde. Chromatin is then cut using a restriction enzyme which is chosen depending on the level of resolution required as different restriction enzymes have different frequencies of cuts along the genome and produce different fragment lengths [162]. The sticky ends of DNA are then re-ligated in extremely dilute conditions to avoid ligation between DNA strands that are not interacting and favor intra-molecular ligation. The resulting structure contains parts of both the DNA fragments that were cross-linked. Finally, these hybrid DNA structures are quantified to generate the number of interactions between DNA fragments [163–165]. However, 3C falls short in genome-wide data analyses. Thus, an adaptation of the technique called Hi-C where chromosomal interactions on a genome-wide scale can be studied simultaneously is the data we have used in order to identify interactions (**Paper III**).

Hi-C was introduced to quantify DNA interactions in an "all vs. all" fashion, i.e. genome-scale chromosomal interactions (Figure 9) [166]. There is only one additional step (after digestion) in Hi-C compared to 3C, in which the sticky DNA ends are labeled with biotin-labeled nucleotides before purifying and shearing DNA. The hybrid DNA molecules are purified by pulling down the biotin mark, and ligated in dilute conditions. The resulting library consisting of hybrid DNA is sequenced using paired-end sequencing. The mapping position of each mate in

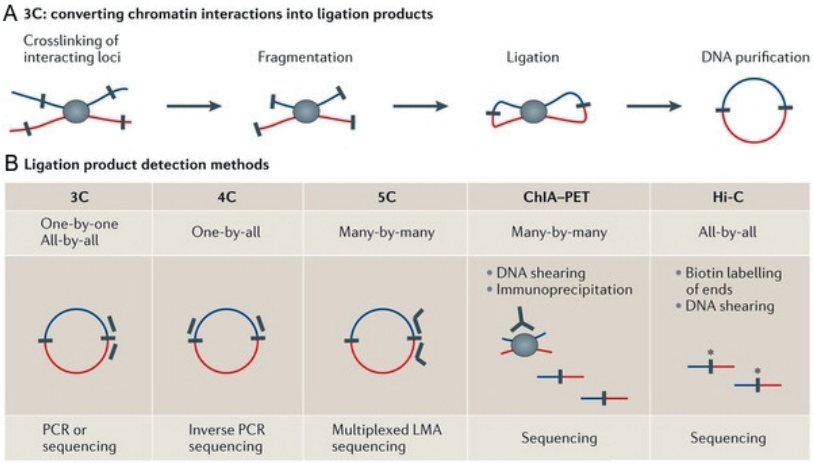

Figure 9: (A) First steps of a 3C protocol. (B) Schematic representation of the principles in various 3C based methods. Figure taken from [163]

the pair is used to generate a genome-wide aggregated contact matrix. Using Hi-C, interactions in the human genome have been mapped at a resolution of 100 kb [166] to 40 kb [119] and more recently 1 kb [167]. Hi-C has also been performed on single cells, providing maps for individual cells instead of cell populations [168]. Thus, Hi-C can provide chromatin-interaction information on a genome-wide level.

Identification of genomic interactions where the same protein is (in)directly interacting with both loci requires **ChIA-PET** [169]. ChIA-PET combines ChIP-seq and 3C to identify regions of DNA bound to the same protein Figure 9 [125]. In other words, it enables identification of protein-mediated DNA-DNA interactions. In the ChIA-PET approach, chromatin is fragmented by sonication after crosslinking and processed for ChIP using an antibody against the protein of interest (similar to ChIP-seq; Figure 8) [125]. Biotinylated DNA linkers added to the ChIPed DNA fragment ends, containing *Mme1* specific restriction sites. The cross-links are reversed; digested with *Mme1*, and biotin-containing fragments are sequenced by paired-end sequencing [164, 170]. Two ligation products are generated by ChiA-PET. One is self-ligations caused by self-circularization ligation

of the same DNA fragment which leads to the reads being mapped close to each other in the reference genome. These reads can be used as anchors of the interacting chromatin segments. The second product is inter-ligations which are pairs of reads further away from each other. These are quantified to generate interaction frequencies between the anchors [120]. Thus ChIA-PET, unlike other 3C-based methods, identifies interactions in DNA where the same protein is bound.

1.4.3 RNA-seq

RNA-sequencing (RNA-seq) uses the power of sequencing to estimate the total amount of RNA in cells [171]. For RNA-seq, isolated RNA is converted to a library of cDNA fragments with adapters and the cDNA library amplified. Since sequencing reads have a read length limit, RNA has to be fragmented into bits of $\sim$200-500 bp. Fragmentation is done by either RNA fragmentation or cDNA fragmentation. Each method has its own bias. Reads from RNA fragmentation are depleted for the ends of the transcript whereas cDNA fragmentation is biased towards the identification of 3' ends of transcripts. The resulting molecules can be sequenced using either single-end sequencing or paired-end sequencing. The sequenced reads can be mapped back to a reference genome or be assembled *de novo*, providing a genome-wide map of transcript levels.

A subset of RNAs such as poly-adenylated (polyA) RNA, micro-RNA (miRNA) [172–175] can be quantified using RNA-seq. MiRNA is a class of RNA which are 21-25 nucleotides long and play a role in gene regulation by silencing genes through RNA-induced silencing complex which contains Dicer and other associated proteins [176–179]. In order to sequence miRNA, RNA is isolated based on size by using a size exclusion gel, or using size selection magnetic beads. On the other hand, mRNA require different steps prior to sequencing. A chain of 100-250 adenines (polyA) are added to the 3' end of mRNA after transcription to make it more stable and prevent degradation [180]. PolyA RNA are mature mRNA which are transported from the nucleus to the cytoplasm to be translated. PolyA RNA can be captured using magnetic beads coated with polyT oligonucleotides [173, 181]. Separating mRNA with polyA tails from total RNA can provide expression levels for coding genes and non-coding transcriptome separately. It has been noted that only about one fifth of transcription in the human gen-

ome is for protein-coding genes [182]. Studies have shown regulation of ncRNAs during development [183, 184]. They also display cell-type specific expression patterns [185, 186] and have shown to be associated with disease [187, 188]. In our study, we have generated a dataset of total RNA as it can provide information on protein-coding genes as well as non-coding RNA which might play an important role in gene regulation.

It should be noted that RNA-seq is conceptually similar to Expressed Tag Sequencing (EST) (an application of Sanger sequencing). However, the fragments to be sequenced have to be $\sim$200-500 bp in length. This generates complications and biases specific to RNA-seq in regard to the length of the transcripts. Nonetheless, it is a massive step forwards from EST and microarray techniques. RNA-seq provides single base resolution without the background noise of microarrays caused by incorrect hybridization while also not requiring full knowledge of a reference genome. Using RNA-seq to quantify the raw sequences of the transcripts present in the cells allows to study not only levels of transcripts but also alternative splicing, mutations, fusion transcripts [189]. Thus, RNA-seq in conjunction with ChIP-seq and 3C-based methods can provide decryption of the functionality of the genome; these approaches have been exploited in this thesis.

1.5 High-throughput sequencing: methods and analysis

Each of the applications of the sequencing methods described in section 1.4, such as ChIP-seq, Hi-C, ChIA-PET and RNA-seq, results in the generation of different datasets with different properties and characteristics. Thus, a multitude of tools and statistical methods have been developed to analyze results from such data, using separate statistical methods for each data-type. In this section, the main analysis methods and tools for the different technologies are assessed.

1.5.1 Sequence alignment to a reference genome

Sequencing data provides a list of short sequences whose original position in the genome is unknown. The most probable location of the sequence can be iden-

tified by mapping the sequence back to the appropriate reference genome. This would be a straightforward task if the entire genomic sequence was unique. However, there are numerous regions with similar sequences such as repeat regions, transposons, and gene paralogs. In addition, single nucleotide polymorphisms (SNPs), where only one base pair is mutated, leads to variations between the sequenced data and the reference genome. There is also a possibility of errors in identifying bases during the sequencing process. However, the sequences are often accompanied by scores providing accuracy estimates for each base that is sequenced. Correspondingly, sequencing aligners have been developed to rapidly and accurately map HTS data to a reference genome, taking the accuracy scores into account.

Popular aligners used for mapping sequencing data are typically based on the Burrows-Wheeler Transform (BWT) [190] and the FM (Ferragina-Manzini) index [191]. The BWT algorithm was originally developed to compress data, by constructing a reversible permutation of the character sequences. The resulting BWT allows for fast lookup of sequences in the database. The FM index identifies exact matches and further builds inexact alignments supported by the exact matches found using a suffix tree. The advantage provided is that substrings of the sequence with multiple copies in the reference genome need to be aligned only once and later collapse into one path as the algorithm traverses the tree. There are more than 60 sequence mappers developed [192] of which Bowtie2 [193] and BWA [194] are popular choices based on the FM index because they are fast and computationally efficient. Sequence alignment is the basic step used in processing of HTS data and is followed by specific analyses.

1.5.2 ChIP-seq analysis

ChIP-seq has been used to identify binding patterns of TFs and histone PTMs. TFs and most histone modifications like H3K4me3 (Figure 10, red) have a narrow chromatin-binding region, whereas H3K27me3 and H3K36me3 (Figure 10, green) are usually wider [195–197]. Enriched regions are identified using peak callers. The most common peak caller used for TFs and histone modifications is **MACS** [198]. MACS takes advantage of observed bimodal enrichment patterns by empirically modeling the shift size. It takes into account that there are

local biases in the genome by modeling the local background using a dynamic Poisson distribution. It identifies peaks with a P-value lower than 10^{-5} (default) which can be tweaked to a lower value for sharp binding patterns and a higher value for broader binding patterns. MACS does not require a control (commonly DNA input) sample as a background. However, it is advisable to have a control for ChIP-seq as data obtained depends highly on sequencing depth, antibody specificity, and variations of enrichment in the cell population. MACS also empirically estimates the false discovery rate (FDR) if a control is provided. To this end, MACS first identifies ChIP peaks relative to the control or background signal, and reverses this procedure by identifying peaks in the control, relative to the observed data. The division of the number of control peaks by the number of ChIP peaks provides an estimate of the empirical FDR. As a result, peaks are generated in ChIP-seq data to identify significantly interacting regions.

Other proteins mapped by ChIP-seq may reveal wide domains of enrichment, such as nuclear lamins (Figure 10) [93]. Similarly, H3K9me3 and H3K27me3 also typically generate a diffuse enrichment pattern along the genome; of note, we found that H2BS112GlcNAc also produces a diffuse enrichment pattern; see **Paper II**. Due to low signal-to-noise ratio, it is not feasible to use traditional peak callers like MACS to identify significantly binding regions for broad binding patterns. To determine binding domains, algorithms like Sicer [199], BroadPeak [200], RSEG [201] and EDD [93] have been developed.

EDD, developed in our laboratory, provides a better overall match with enrichment patterns for ChIP-seq data with large domains. This is because EDD is highly robust against local variations and works with the size of the domains rather than signal-strength. Binning the genome is a key first step for EDD. Binning is performed by dividing the length of the genome into equal sizes. For example, a sequence of 100 kb can be divided into 10 bins of 10 kb. EDD divides the genome in bins of equal size and calculates the smallest bin size that would provide the most information by estimating the signal in the bin using the Agresti-Coull method [202]. The signal in the bins is displayed as the ratio of ChIP reads over input reads where input is normalized by the ratio of sequencing depth difference between the samples. Thus, noisy lamin A/C data (Figure 10, gray) is transformed into a clear positive or negative signal (Figure 10, black) for each bin. EDD generates peaks by scoring consecutive bins where non-informative and depleted bins

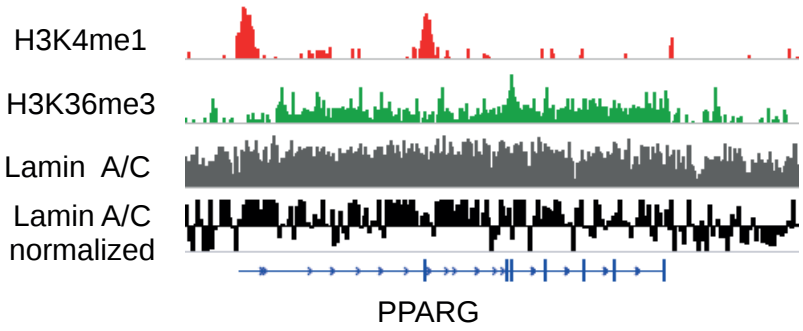

Figure 10: Different types of enrichment profiles generated by the mapping of ChIP-seq reads. H3K4me1 produces narrow peaks and H3K36me3 produces wider peaks. These can be analyzed using traditional peak callers. Lamin A/C has an extremely diffuse chromatin binding pattern. Normalized data allows to view enrichments and depletions in the sample over background levels.

are assigned a gap penalty. Domains are detected by using a linear algorithm for identifying maximal scoring subsequences [203] and finally assigning P-values using Monte Carlo trials [204]. Thus, using EDD as a domain caller for ChIP-seq data with diffuse binding patterns like lamin A/C, H2BS112GlcNAc provides a solid base for downstream analyses for such data sets as shown by Lund et al. [93].

1.5.3 Modeling of chromatin states

Segmentation of genomic data helps differentiate the genomic regions based on their functionality. As mentioned earlier (section 1.1.1), regions of the genome are defined based on their functionality, such as promoters, enhancers, exons, etc. and these regions have been found to house specific histone modifications. There has been an exponential increase in the availability of epigenetic data in recent years with the large scale sequencing projects carried out by consortia such as ENCODE Consortium [131], Epigenomics Roadmap [205], and Fantom [206]. Thus, it has become increasingly common to study combinations of epigenetic marks. These combinations may encode distinct biological functions [207]; however, the functional significance of combinations of epigenetic marks are mostly unknown. Studying combinations of histone modifications in tandem by analyzing multiple

data-sets with distinct sets of signals and peaks is difficult to parse. Thus, in addition to aiding interpretation, studying various combinations of epigenetic modifications based on their segmentation along the genome gives a deeper understanding of the co-occurrence of chromatin marks on the genome. ChromaSig [208], Segway [48] and ChromHMM [209] are the three main methods performing segmentation based on epigenetic marks.

ChromaSig, released in 2008, works in two steps. First, it identifies 2 kb loci which are highly enriched in chromatin marks. In order to take chromatin signatures in the vicinity into account, a 7 kb window around the enriched 2 kb loci is searched to generate a signature motif pattern of 4 kb. A 4 kb length ensures that the motif covers at least 75% of the enriched loci. Second, ChromaSig clusters, aligns and orients the enriched loci identified in the first step based on the Euclidean distance between the chromatin modifications in the motifs. This method was a breakthrough in genomic segmentation. However, unlike later algorithms, ChromaSig does not perform genome-wide segmentation.

ChromHMM segments the genome using a multivariate Hidden Markov Model (HMM) [209]. ChromHMM finds a local optimum of the parameter values using a variant of the standard expectation-maximization based Baum-Welch algorithm. In this case, the algorithm completes one iteration over the chromosomes for the dataset, and applies an incremental expectation-maximization procedure so that parameter estimates are incorporated rapidly. The algorithm, by default performs 200 iterations. However, it should be noted that often, with numerous data-sets, the algorithm will not demonstrate convergence in 200 iterations and hence, more iterations should be performed if required. The model can be explained as illustrated in Figure 11. In this example (Figure 11A), there are three states (P = promoter, G = gene body, B = blank) which can be identified using H3K4me3 and H3K36me3. A state would, for example, be a promoter state if H3K4me3 has an emission value of 1 and H3K36me3 has an emission value of 0. The transition probabilities for going from one state to another are displayed by the arrows. A promoter state has the highest transition probability (0.8) to a gene body and a low transition probability to a promoter state (0.1) or a blank state (0.1) whereas a blank state would have almost no chance (0.01) of transitioning to a gene body. In the case of calling states using ChromHMM, the sequence of observations is known (where the histone PTMs are located), and the sequence of states (P, G, B;

see above) is unknown. Thus, using the forward-backward algorithm can provide the posterior probability distribution for each hidden state interval, conditional on the observed data. In order to identify states in the genomic sequence, the sequence can be divided into bins and the most likely hidden state, for each bin, can be inferred based on the trained model. In Figure 11B, the sequence starts with a B since there is no signal for either histone PTMs. The probability for the next state, which is B, is provided by the forward-backward algorithm. The forward-backward algorithm first calculates the transition and emission probabilities in a forward and backward pass. A smoothed value is generated by combining the two probabilities. Thus, two states which are rarely observed consecutively, can still be found next to each other using the forward-backward algorithm (Viterbi algorithm can be used instead to identify the most probable sequence of states). In the first step, the algorithm computes a set of forward probabilities to end up in state B given the observations in the previous states. In the second step, the algorithm computes a set of backward probabilities which would determine the probability of observing the remaining states given any starting state. Thus, the two probabilities can be combined to obtain the distribution over any specific bin in the sequence given the entire observation sequence. Continuing in the sequence, the third bin has H3K4me3 and a low signal from H3K36me3. The transition probabilities and emission probabilities, in this case, dictate the state transition to P. In this manner, ChromHMM segments the genome using multivariate HMMs which are graphical probabilistic models that model multiple 'observed' inputs generated by unobserved 'hidden' states, using transitions between hidden states to model spatial relationships [209].

The number of states to be modeled by ChromHMM is decided by the user. Given a study with 7 tracks of histone modification data, there is theoretically a chance of finding 7! (factorial 7) combinations of present/absent signal for each bin in the genome which would mean a total of 5040 possible states. However, that would create too many states to analyze. Additionally, it would also make more biological sense to only have a combination of certain genomic elements, such as e.g. H3K4me1 with H3K27ac which mark enhancers. It is also important to clearly consider what the end-result of the state-calling should be. If the user wants to easily interpret the data, few (and easily interpretable) states should be chosen. So, it is important to try a varied number of state calls which would

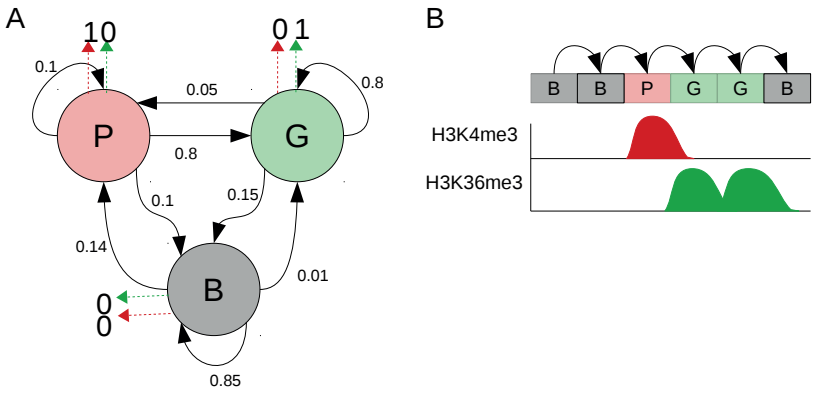

Figure 11: Illustration of HMM. (A) Three states (Promoter = P, Gene body = G, Blank = B) have their own emission probabilities (red, green dotted arrows) and each state has a transition probability (black arrows) to move from one state to another. The emission and transition probabilities dictate the state call together. (B) Generation of a chain of states based on the observations and probabilities from the trained model.

explain the datasets concisely and precisely. The authors of ChromHMM [209] used the Bayesian Information Criterion (BIC) to identify the optimal number of states to be included in the model. BIC, in simple words, penalizes the inclusion of extra model parameters that do not increase the likelihood accordingly. The optimal number of states is therefore given by the model with the lowest possible BIC score. The model generated can be pruned to eliminate states such that the remaining states have the least distance from their closest emission vector in the remaining states. For example, ChromHMM can be used to generate a 15 state model using seven datasets as shown in Figure 12. For example, state 6 (dark green) is marked by only H3K36me3, state 13 (yellow) is marked by a combination of H3K27ac, H3K4me2, H3K4me3 whereas state 7 (black) is devoid of the marks being studied here. After the emissions for the 15 states have been generated, the user can annotate the states manually using prior knowledge on the functions of the chromatin marks. Thus, identifying the optimum number of states is a compromise between resolution and diversification of the data, and is dependent on the overall goal of the study.

Segway and ChromHMM are very similar in their final output and method.

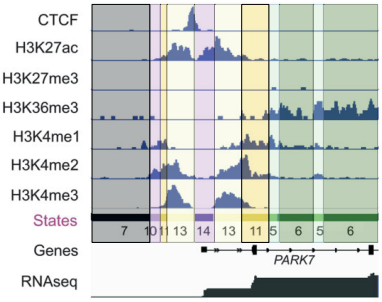

Figure 12: ChromHMM segmentation. Epigenetic information is condensed to allow visualization and further down-stream analyses on multiple epigenetic marks at a single location.

Segway uses a Dynamic Bayesian Network (DBN) to model the complex hidden relationships which explain the observed data sampled at regular intervals of the genome as an axis [48]. A DBN can represent a standard HMM through a random variable for the hidden state of the HMM along with an observed random variable for the observations. However, a DBN can incorporate complex relationships among variables. A DBN models the interrelationship between multiple hidden variables without the need to flatten them into one variable. Segway simultaneously segments and clusters genomic data. The authors of Segway have shown that this algorithm is capable of "rediscovering" annotated regions of the genome using DNase-seq data, FAIRE-seq and ENCODE data by unsupervised training of a portion of the human genome [48]. They also compared their findings with ChromHMM and identified that Segway generated a finer-grained segmentation in comparison to the former. ChromHMM segmentation with the same dataset generated segments with a mean length of 4,862 bp and a median of 800 bp whereas Segway had a mean length of 168 bp and a median of 124 bp. Thus, Segway has a better ability to detect elements at a sub-nucleosomal resolution [48].

1.5.4 Hi-C analysis

Hi-C provides high resolution maps for genome-wide chromosomal interaction. As with ChIP-seq, Hi-C analysis begins with mapping of sequencing data to a reference genome. Hi-C always requires paired-end sequencing. Thus, mapping

both mates of each paired sequence can result in (i) none of the mates in the pair being uniquely mapped (ii) only one mate of the pair being uniquely mapped (iii) both mates mapping to the same bin (iv) both the mates in a pair being mapped to separate bins. In the first case, the pair is discarded. In the second case, the mate that mapped can be saved to be included in downstream analysis, for example, avoiding artifacts resulting from excluding reads that have frequent interactions with repeat regions from one end. The third case is caused due to the presence of dangling-ends caused due to self-circularization or self-ligation and these reads should be discarded. There is also a bias caused by PCR amplification and all reads that map at the same location multiple times should be removed [210]. All reads from the fourth case should be saved for identifying significant interactions between chromatin segments [210].

In one of the first studies to identify significant Hi-C interactions, the authors sought to infer interactions that were more frequent than the background signal [211]. The authors treated the inter-chromosomal and intra-chromosomal reads separately to identify significant interactions in each set. A uniform probability model was assumed for interactions between pairs of restriction fragments. The probability of observing any particular interaction was calculated as a ratio of one divided by the total number of possible inter-chromosomal pairs of the restriction fragments. The probability of observing a particular number of inter-chromosomal interactions was given by the binomial distribution. However, for intra-chromosomal interactions, the genomic distance had to be accounted for since chromosomes act as polymers and have a higher probability for random contacts with shorter genomic distance. This was corrected by the authors by grouping the contacts in discrete 5 kb bins and the probability within each bin was calculated separately. Thus, as noted by the authors, the probability of the contacts is conditioned by the genomic distance.

The above method was refined in a follow-up study which focused on identifying significant interactions at a genomic scale of $\sim$50 kb to 1 Mb [212]. The authors modeled a combination of the random interaction due to genomic distance and other biases in Hi-C data. These biases include a non-uniform distribution in restriction fragments in the genome due to different frequency of restriction sites in each bin. Biases also include GC-content, fragment length and mappability [213]. The bias was incorporated by using an iterative correction and

eigenvector decomposition model [210] in conjunction with a monotonic spline fitting procedure. The authors excluded bins with a very high or low bias when generating P-values.

1.5.5 ChiA-PET analysis

ChiA-PET provides interaction maps within the genome, mediated by a selected protein of interest that has been immunoprecipitated. Mapping paired-end ChiA-PET data results in two types of ligation products: self-ligations and inter-ligations (see section 4.2). Since ChiA-PET does not map contacts for the entire genome, it is required to first identify the sites (anchors) where interactions occur. For this, the inter-ligations can be used in a similar fashion as ChIP-seq data (see section 1.5.2), to detect peaks (anchors) [120, 125]. Pairs with one mate mapping to each of these two anchors are counted to identify the raw signal between each pair. ChiaSig [214] can be used to identify significant interactions between such anchors across the genome. ChiaSig was the first method that took the genomic distance between the anchors into account, when inferring the significance of the interactions. ChiaSig extends the hypergeometric distribution used by previous methods [120] and replaces it with the non-central hypergeometric distribution, and has been proved to identify relevant significant interactions and averting the over-estimation of shorter interactions. Thus, by using defined regions as anchors in conjunction with ChiaSig, it is possible to identify significant interactions between only regions associated by a given protein or histone PTM of interest.

1.5.6 Predicting chromatin interactions

Since generating Hi-C and ChiA-PET data-sets is expensive and exhaustive, methods are being developed to predict interactions in the genome by modeling known information instead of producing Hi-C datasets for each cell type. As opposed to the prediction of putative enhancers (by p300, H3K4me1, H3K27ac), there is no consensus predictor of interaction between promoters and enhancers. Several unsupervised and supervised methods have been proposed to identify interacting P-E pairs.

One of the most basic **unsupervised** methods for predicting P-E interactions is to simply select the nearest enhancer to the promoter [215, 216]. However, only

a fraction ($\sim$40%) of enhancers interacts with the nearest promoter [215, 216]. Thus, selecting P-E pairs only in the same TAD has provided better results [119, 149]. One variant uses sequence co-conservation for the promoter-enhancer pair [217,218]. Another approach is to select promoters within a certain distance from each enhancer and identify interactions based on activity correlations for DNase I hypersensitivity sites [219]. The authors also corroborate the findings with 5C and discovered that roughly only half of the predicted promoter-enhancer pairs were markedly enriched in long-range interactions but not necessarily significant. Ernst et al. developed a method that uses correlation between promoter and enhancer states in conjunction with genomic distance (minimum distance of 5 kb; maximum distance of 125 kb). This method incorporates both genomic distance and correlation values in a logistic regression classifier to identify 'real' P-E interactions [47].

Recently, PreSTIGE was developed to identify cell-type specific P-E interactions [220]. PreSTIGE first identifies enhancers defined by H3K4me1 and provides specificity scores based on Shanon's entropy [221], and cell-type specific enhancers are considered active. The step involves pairing the enhancer to a promoter which is closer to the enhancer site than the closest CTCF site, or 100 kb at the farthest. Considering the complications and chromatin marks required for predicting P-E interactions, it is imperative to use more features and non-linear models for predicting P-E interactions.

Supervised methods have been developed to predict P-E interactions. Random Forest (RF) classification is a popular method employed for this purpose. It is better to use tree-based methods as opposed to linear regression methods since there is not always a linear function that can explain the relationship between genomic datasets [222]. IM-PET is an example of an algorithm developed to use RF to predict P-E interactions [216]. On average, 36,823 P-E interactions from 12 cell-types were identified. The authors used promoter activity level, TF binding probability based on binding motif, sequence and synteny conservation information, and distance between promoters and enhancers as the features for generating the RF model. Another algorithm, RIPPLE, also employs RF to identify P-E interactions [222]. The authors selected the set of features that worked best across four cell types which were CTCF, cohesin (RAD21), H3K4me2, H2K27ac, H3K9ac, H3K36me3, H4K20me1, H3K27me3, DNase I, and TBP (a TF). The authors

cross-checked the predicted genome-wide maps using Hi-C data and identified that high confidence interactions (in the 90% percentile) had significantly more contact counts in the Hi-C than low confidence intervals (in 10% percentile). This corroborates the prowess of RF in predicting genomic interactions.

RF is a tree-based method which creates multiple trees from many random subsets of the data and tests the robustness of the model on another subset of the data, thus providing information on the importance of each of the features in the dataset and error rate of prediction using those features. The process of a tree generation is illustrated in Figure 13. For example, the full set consists of P-E pairs with two possible outcomes (i) significant interactions and (ii) non-interactions. Each pair, regardless of the outcome, is quantified by a set of features as shown by the shapes. A tree is built by taking a random subset from the full set of observations and dividing it into a training set (2/3 of the subsample) and a test set (1/3 of the subsample, known as "Out-of-Bag" (OOB)). Simultaneously, a random subset is generated from the list of describing features as well. The features act as the nodes of the tree. The outcomes in the training set are split based on the best classifier from the first set of random features and the remaining outcomes are subsequently split using another feature from another random set of features and so on until the terminal nodes (leaves) of the tree contain outcomes of only one value. Once a tree is generated from the training set, the P-E pairs from the test set traverse down the tree based on the values of their features, and their outcome is predicted and subsequently compared with its known outcome. The percentage of false predictions provides the OOB error rate. This process of training and testing is iterated multiple times leading to the formation of a forest and an overall OOB error rate. Eventually, through the randomization process, each pair will have both helped construct a tree and test a tree. Once a model has been generated, the RF can predict the outcome for P-E pairs with unknown outcomes (unknown set). These pairs will however need to have a value for the features used to develop the RF. Each P-E pair from the unknown set will traverse down all the trees generated in the model. The predicted outcome for each P-E pair from each tree is counted as a vote. The ratio of the votes determines the outcome of the P-E pair. For example, if for one P-E pair, there are 300 votes for significant interaction and 100 votes for no interaction. Then the outcome is significant interaction for the P-E pair with a classification probability of 0.75. Thus, random forests can be

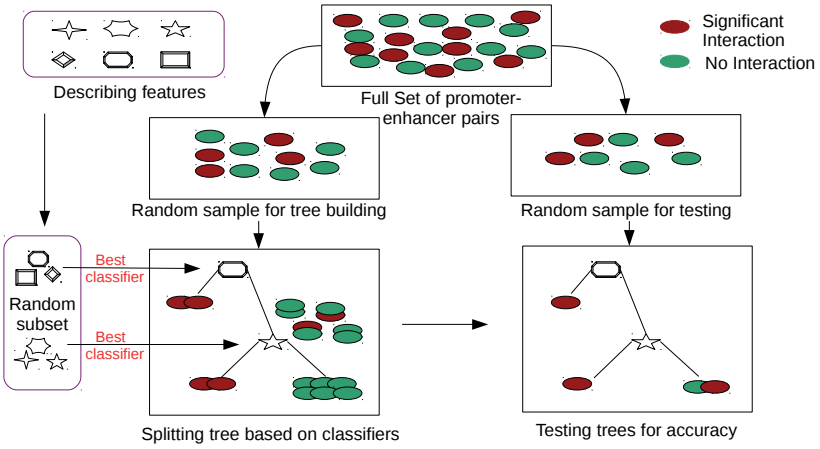

Figure 13: Random forest generation. Random samples are taken from the full set of known data to generate a random forest model and test it on a subset to quantify the accuracy of the model.

used to predict the outcome of results based on features that are relevant to the outcomes in a robust way by generating results from thousands of permutations.

1.5.7 RNA-seq analysis

RNA-seq is an HTS-based tool for estimating the relative abundance of transcripts in a cell or a population of cells. Expression levels in RNA-seq are typically measured in either **r**eads **p**er **k**b per **m**illion base pairs mapped (RPKM), or **f**ragments **p**er **k**ilobase per **m**illion base pairs mapped (FPKM) [173]. RPKM is calculated for single end RNA-seq as there is only one read that provides information about the transcript. FPKM is calculated for paired-end RNA-seq when two mates of a pair provide information about the sequenced transcript as in this case reads are obtained from both ends of one fragment. FPKM is calculated by counting the number of reads mapping to a gene and then dividing that count by the length of the transcript in kb. Expression levels are normalized by dividing them by the total sequencing depth in Mb. This scales the expression levels of genes irrespective of sequencing depth. For example, FPKM of a transcript that is 10 kb in length with 1000 fragments mapping on the transcript at a sequencing

depth of 50 Mb would be calculated as follows:

FPK (Fragments Per kb (of transcript)) = 1000 / 10 = 100

FPKM (FPK per Million (of mapped fragments)) = 100 / 50 = 2

There are four major steps typically used for analyzing RNA-seq data.

(i) Mapping of data to a reference genome. Genome alignment from ChIP-seq data depend on the BWT and FM index for rapid alignment. However, such methods are not optimal for mapping RNA-seq data, since the data stem from RNA transcripts instead of genomic DNA. Due to splicing, RNA editing, and the presence of multiple RNA isoforms, mapping the data back to the genomic sequence is not trivial. This process can be complicated by incomplete knowledge of the reference genome (and transcriptome) and aberrant splicing events in mutated cells [223]. Thus, a mapping algorithm for RNA-seq should be able to align reads across splice junctions and handle paired-end reads from long-range splices in a reasonable time-frame. **TopHat2** is one of the most popular RNA-seq mapping algorithms which satisfy these criteria [224]. TopHat2 first maps all reads to a transcriptome with known junctions. The reads that do not map are then mapped to the reference genome to identify unannotated exons. This step creates two lists. One with reads spanning a single unannotated exon which are mapped and multi-exon spanning reads which are unmapped. The unmapped reads are then split into smaller segments which are realigned to the reference genome and reads for which both the left and right segments are mapped, the most likely splice junctions are identified. Further, to align reads that do not map in the previous step, flanks of exons are concatenated to form a new transcriptome and unmapped reads are aligned again. Finally, TopHat2 realigns multi-mapped reads based on the new transcriptome and splice junctions, and reports only the most likely locations for the mapped reads. Thus, TopHat2 manages to map as many reads as possible from the sequenced sample ensuring minimal loss of information.

(ii) Transcriptome reconstruction and quantification. Mapping the data only informs the location of the sequences in the reference genome. These sequences need to be allocated to their respective known transcripts to estimate the abundance of various transcripts of a gene while simultaneously estimating alternative isoforms which is performed by algorithms such as Cufflinks [225]. Cufflinks first identifies 'incompatible' fragment pairs which most likely originate from distinct spliced isoforms. An overlap graph is generated by connecting

'compatible' fragments that overlap in the genome with each fragment having a node in the graph. For example (Figure 14A), the red, blue and yellow fragments might be separate isoforms and the black fragments could originate from either of the three isoforms as one end in each of these fragments does not overlap another fragment in the sample. The minimal path covering the mutually compatible fragments for the three isoforms is shown in Figure 14B. This is in concordance with Dilworth's theorem [226] where (in RNA-seq context) all fragments can be explained when the number of isoforms is equal to the number of mutually incompatible fragments. The previous steps are done in small bundles to be computationally fast and efficient. The results are then concatenated to identify abundance for each of the transcripts. This is done by matching fragments to the possible transcripts. For example (Figure 14C), the black fragments can be part of all three isoforms whereas the violet fragment can be in either the red or the blue isoform. It can only be assigned to one isoform as assigning to multiple isoforms would imply different cDNA lengths for the same fragment. It would be improbable for the violet fragment to originate from the red isoform due to its larger length. Thus, it would be assigned to the blue isoform. In order to confirm this, Cufflinks assigns the maximum-likelihood abundance to each isoform which can best explain the mapped fragments as illustrated in the pie chart (Figure 14D).

(iii) Differential gene expression analysis. Identifying differentially expressed genes is not a necessary step for analyzing RNA-seq data and is only necessary to compare the expression levels between two or more treatments or differentiation stages of cells. Algorithms used for identifying differentially expressed genes between data sets have been scrutinized using simulated [227] and real data [228] to assess their efficiency and accuracy to produce a list of differentially expressed genes. Data from RNA-seq experiments generate a skewed distribution with the majority of the genes having expression at the lower end of the scale and thus making it difficult to use parametric methods to identify significantly differentially expressed genes from raw counts. Methods developed for this purpose are mainly based on negative binomial models (edgeR [229], DEseq [230]), non-parametric approaches (SAMseq [231]), and transcript-based methods (Cuffdiff2 [232]). **edgeR** uses an empirical Bayes procedure to moderate over-dispersion across genes by borrowing information from other genes. The data among samples is normalized using a Trimmed mean of M values method

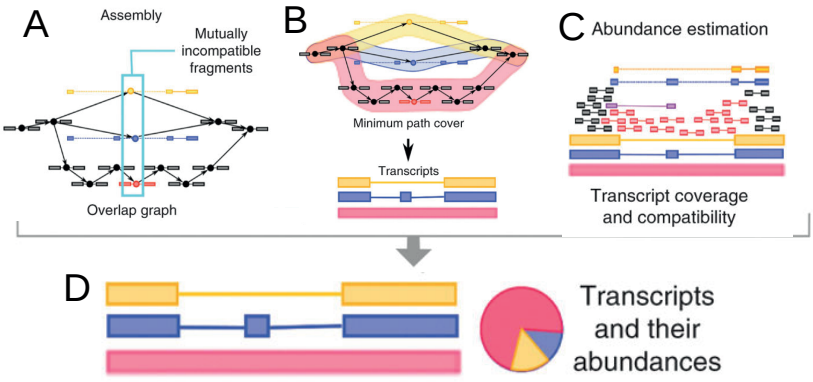

Figure 14: Overview of Cufflinks. (A) Possible isoforms are identified from the mapped fragments based on overlap. (B) A graph is generated and transcripts identified based on the shortest path. (C) Abundance for each transcript is measured based on the most likely transcript a fragment originated from. (D) Transcripts are quantified by the maximum-likelihood abundance. Figure adapted from [225]

[233]. It uses an exact test adapted from Fisher's exact test to incorporate over-dispersed data and uses Benjamini-Hochberg method [234] to generate FDR values. **SAMSeq** uses the non-parametric Wilcox rank test in conjunction with a resampling procedure (accounts for variable sequencing depth) followed by a permutation-based method to generate FDR values. **Cuffdiff2** uses a beta negative binomial model for fragment counts. This controls variability and ambiguous read mapping as it estimates expression at a higher resolution of transcript level. The resolution of Cuffdiff2 also enables it to provide differential expression at promoter, isoform and gene levels. It normalizes data to the sequencing depth and uses Benjamini-Hochberg method to generate FDR values. There are pros and cons for each method as identified by different comparative studies and the consensus remains divided [227, 228, 235]. Statistical tests based on negative binomial methods provide the highest number of differential genes whereas transcript based methods tend to identify the fewest. Thus, multiple methods must be tested before selecting an algorithm to identify differentially expressed genes.

(iv) Plotting and subsequent downstream analysis. This is the least developed step for RNA-seq analysis for in-depth cross data analysis. Most of the

packages for RNA-seq analysis do produce plots for basic analysis. **The Tuxedo protocol** [236] has provided a dedicated bioconductor [237] package to generate plots, albeit, only for data generated from Cuffdiff. However, plots can also easily be generated by using other packages in R [238, 239] like ggplot2 [240]. Another important way of showing RNA-seq (and/or ChIP-seq, and/or Hi-C) data is with the use of genomic browsers such as Integrative Genomics Viewer (IGV) [241]. Loading data from RNA-seq analysis into IGV allows visualizing the difference in expression levels of genes between genes and/or loci along the genome. One can also add TF and histone PTM ChIP-seq data to obtain a graphical representation of the chromatin environment of a given gene expression output (see **Papers I** and **II**). Other downstream analyses can be performed based on transcriptomic data, such as Gene Ontotology (GO analysis), Principal Component Analysis (PCA), clustering based on expression values as we have done in **Paper I**.

1.6 Summary

The advent of high-throughput sequencing technology has led to the production of a wealth of data which have given totally new insights into our understanding of genome function. However, there is a huge gap in the knowledge of how epigenetic modifications, protein-DNA interactions and interactions between genomic elements, e.g. between an enhancer and its cognate promoter, transition in the course of development and cell differentiation to lead to the multitude of cell types that make up an organism. In the doctoral work presented here, we have largely relied on the differentiation of human adipose tissue stromal cells (ASCs) into adipocytes [242] to unveil temporal changes in gene expression and associated chromatin modifications. This adipogenic differentiation system has been established earlier in our laboratory [243] and is described in **Paper I**.

We then substantially establish that adipogenic differentiation is associated with massive spatial reorganization of lamin A/C across the genome, through dynamic changes in interactions of chromatin with nuclear lamins. We further identify an epigenetic pre-patterning of *de novo* nuclear lamin-chromatin interactions in this process, involving genomic domains of GlcNAcylated histone H2B (**Paper II**). Finally, we propose a method to study interactions between promoters and enhancers where genome-wide interaction data are unavailable, and

predict promoter-enhancer interactions taking place during adipogenic differentiation (**Paper III**).

Chapter 2

Aims of the study

Chromatin remodeling is critical for the correct programing of developmental gene expression. Recent work has provided a dynamic view of post-translational histone modifications during cell differentiation; however there has been little insight on the evolution of combinatorial genome-wide patterns of chromatin marks, excluding an essential aspect of developmental gene regulation. Similarly, very few studies have addressed the nature of global changes in chromatin organization during differentiation. Using an *in vitro* adipogenic differentiation model, this thesis addresses relationships between chromatin organization and developmental gene expression patterns.

The aims of this study were therefore to:

- Identify chromatin states and their temporal relationship to gene expression patterns during differentiation of human adipose tissue stem cells
- Determine whether adipogenic induction affects the genome-wide distribution of the nutrient-responsive S112GlcNAc modification on histone H2B, and bioinformatically characterize H2BS112GlcNAc domains (GADs)
- Map changes in the genome-wide association of chromatin with nuclear lamin A/C during adipogenic differentiation, and characterize these associations
- Assess the relationship between changes in lamin A/C-chromatin interactions and GADs driven by adipogenic differentiation, and their relationship

to gene expression

- Develop a predictive modeling technique to infer significant and dynamic promoter-enhancer interactions in differentiating adipose tissue stem cells, for which no chromosomal interaction data are available

Collectively, this work provides significant insights on our understanding of spatial genome organization in a dynamic differentiation context. It also provides a new approach to inferring physical interactions between gene regulatory elements.

Chapter 3

Summary of publications

3.1 Paper I: A hyper-dynamic nature of bivalent promoter states underlies coordinated developmental gene expression modules

(Shah et al., 2015 BMC Genomics 15, 1186)

Chromatin remodeling is crucial for proper programing of developmental gene expression. Recent work provides a dynamic view of post-translational histone modifications during differentiation; however there is little insight on the evolution of combinatorial genome-wide patterns of chromatin marks, excluding an essential aspect of developmental gene regulation. We report here a 15-chromatin state Hidden Markov Model which describes changes in chromatin signatures in relation to transcription profiles during differentiation of human pre-adipocytes into adipocytes. We identify nineteen modules of gene expression reflecting multiple waves of transcriptional up- and down-regulation which characterize adipogenic differentiation. From our model, we developed chromatin state matrices fitting each of these transcription modules to show how the complexity and dynamic nature of chromatin signatures relate to expression patterns. Spatial relationships between chromatin states underlie a high-order chromatin organization in differentiating adipocytes. We show the importance of gene expression level in generating diversity in chromatin signatures, and show that the hyper-dynamic nature of

H3K4me2/H3K27me3-marked 'bivalent' promoter states underlies many of the gene expression patterns associated with adipogenic differentiation. Our results reveal the highly dynamic nature of bivalent promoter states within the adipogenic lineage. The data constitute a valuable resource enabling the assessment of possibilities to alter the adipogenic program.

3.2 Paper II: Pre-patterning of differentiation-driven nuclear lamin A/C-interacting chromatin domains by GlcNAcylated H2B

(Rønningen, Shah et al. Genome Res., accepted)

Dynamic interactions of nuclear lamins with chromatin through lamin-associated domains (LADs) contribute to the spatial organization of the genome. We provide here evidence for a pre-patterning of differentiation-driven formation of lamin A/C LADs by domains of histone H2B modified on S112 by the nutrient sensor O-linked N-acetylglucosamine (H2BS112GlcNAc), which we term GADs. We reveal a two-step process of lamin A/C LAD formation during *in vitro* adipogenesis, involving spreading of lamin A/C-chromatin interactions in the transition from progenitor cell proliferation to cell cycle arrest, and genome-scale redistribution of these interactions through a process of LAD exchange within hours of adipogenic induction. Lamin A/C LADs are found both in active and repressive chromatin contexts that can be influenced by differentiation status. We show that de novo formation of adipogenic lamin A/C LADs non-randomly occurs on GADs, which consist of megabase-size intergenic and repressive chromatin domains. Accordingly, whereas pre-differentiation lamin A/C LADs are gene-rich, post-differentiation LADs harbor repressive features reminiscent of lamin B1 LADs identified in other cell types. We find that release of lamin A/C from genes directly involved in glycolysis concurs with their transcriptional upregulation after adipogenic induction, and with downstream elevations in H2BS112GlcNAc levels and O-GlcNAc cycling. Our results reveal an epigenetic pre-patterning of adipogenic LADs by GADs, suggesting a coupling of developmentally regulated lamin A/C-genome interactions to a metabolically-sensitive histone modification.

3.3 Paper III: Inference of promoter-enhancer contacts from epigenomics datasets reveals dynamic interaction during adipogenic differentiation

(Shah et al., manuscript)

Spatial genome conformation is central to the regulation of gene expression. In mammalian genomes, chromatin looping events lead to interactions between promoter and cognate distal enhancers to regulate transcription. Promoter-enhancer (P-E) interactions can be identified genome-wide by high-throughput chromosome conformation capture techniques such as Hi-C and ChIA-PET. We report a predictive modeling technique to infer significant P-E interactions in cell types for which such data are not available. Our methodology relies on training a random forest model from Hi-C and genome-wide epigenomics datasets to predict significant P-E interactions and epigenetic features contributing to these interactions. Accuracy of the predictive value of the model can be modulated by fine-tuning the selection of chromatin features. Our model points to constitutive CTCF binding across many cell types and DNase accessibility as the strongest predictors of significant P-E interactions. Applying our model to infer P-E interactions during differentiation of adipose tissue stem cells identifies dynamic developmentally-linked P-E interactions compatible with the adipogenic transcriptional program. Reliably predicting P-E interactions constitute a valuable tool to investigate the developmental dynamics of genome architecture.

Chapter 4

Discussion

4.1 ChromHMM application to genome-wide data

HTS technology has made multiple datasets available for genome-wide profiling of histone PTMs and chromatin-binding proteins. Genome segmentation methods to identify co-occurrence of chromatin marks have therefore become increasingly popular. These include Segway [48], ChromHMM [209] and ChromaSig [208].

In our work, we have used ChromHMM to segment the genome based on enrichment profiles of seven chromatin marks including 6 histone PTMs (H3K4me1/2/3, H3K27ac/me3, H3K36me3) and CTCF binding. We have not used ChromaSig because it does not provide genome-wide segmentation [208] and therefore was inadequate given the genome-scale nature of our analyses. Segway uses a DBN. Thus, Segway is similar to ChromHMM since a standard HMM can be represented by a DBN. In comparison to ChromHMM however, Segway enables finer genome segmentation without binning the data and handles missing data better [48]. The ability of a DBN to handle missing data is an advantage when the study uses large numbers of epigenetic marks and there is missing data for different datasets. However here, we used data consistent for each adipogenic differentiation time points; thus the question of handling missing data did not arise and we chose not to use Segway. Moreover, we used in our chromatin state analyses relatively few chromatin marks (7); thus generating chromatin states at a finer resolution (e.g. with Segway) would result in excessive segmentation. We opted for a 15 chro-

matin state HMM with a 200 bp fragmentation (**Paper I, II, III**); this provided a robust base to relate with high confidence epigenetic data to lamin A/C binding and H2BGlcNAc enrichment patterns, whose detection required with the larger (1 kb) bins (**Paper II**, section 1.5.2). Thus, information at a sub-nucleosomal level generated by Segway would be noisier and overwhelming in comparison.

Segway generates chromatin states based on a single base-pair resolution as opposed to the binning approach of ChromHMM. Binning size in ChromHMM can be adjusted to below 200 bp [47, 209] to provide deeper resolution. However, ChIP-seq read length is usually at least 50 bp, and chromatin fragmentation for ChIP results in fragments of 200-500 bp, so resolution beyond that size at the genome-scale level is in reality not possible. Thus, the molecular methods do not provide location of the epigenetic mark(s) at base-pair resolution. Instead, aggregation of sequencing reads in a bin provides the most likely location of epigenetic mark(s) in a genomic region. Thus, using 200 bp bins (approximately the size of a nucleosome) in our study seemed most appropriate to identify binding regions for the epigenetic marks considered.

Another difference between Segway and ChromHMM is that Segway uses an inverse hyperbolic sine function to transform the signal values, whereas ChromHMM uses a binary approach to identify the presence or absence of an epigenetic mark in the bin. In fact, another algorithm based on HMM, EpiCSeg [244], also works on signal values for the epigenetic marks. However, EpiCSeg does not take into account the fact that most epigenetic modifications have different binding patterns. For instance, H3K4me3 has a narrow binding pattern whereas H3K36me3 displays broader domains of enrichment. Lamin A/C and H2BGlcNAc have diffuse binding patterns (**Paper II**). This results in varying numbers of signal intensities for each of the epigenetic marks examined, leading to a high probability of false state calls solely due to the lower enrichment level of one of the marks. Through a binary approach, ChromHMM identifies the cut-off for each mark independently using a Poisson distribution, and provides present/absent calls for each mark for each bin. Thus, when using epigenetic data from different sources, it is important to distinguish advantages and disadvantages of signals generated from the marks; this task is performed well by ChromHMM.

Another key difference between Segway and ChromHMM is that Segway uses the Viterbi algorithm to call states whereas ChromHMM uses the forward-

backward algorithm. The advantage of the Viterbi algorithm is that it can identify the probability of a sequence of states. Thus, states with a rare probability of occurring one after another in the genomic sequence cannot be called; in ChromHMM however, if emission and transition parameters are "right" for a given state, the state can be called. This may be a drawback of ChromHMM; however in practice this matters little because the exact sequence of epigenetic marks along the genome is largely unknown since a large number of epigenetic marks have still not been profiled. Thus, despite the possibility to identify rare state sequences, it was not a hindrance to use ChromHMM for segmentation in our study. In summary, it is clear that DBN presents advantages over an HMM, which might be advantageous in settings outside biology. However, in our study, we rationalized that ChromHMM was beneficial over Segway to call chromatin states because it suits the data at hand more appropriately.

4.2 Training a P-E interaction model can be influenced by features chosen

RF is a machine learning method that builds a model based on training data by constructing a multitude of decision trees. In **Paper III**, we used histone PTMs (H3K4me1/2/3, H3K27ac/me3, H3K36me3), CTCF and DNase I hypersensitivity datasets to train a model from six cell types (GM12878, HMEC, HUVEC, IMR90, K562, NHEK); we then used the model to predict P-E interactions in ASCs at four stages of adipogenic differentiation, for which P-E interactions are not known. Importantly, Random Forest can rank the importance of the features fed into the model and returns their prediction power for a P-E interaction.

RF models have previously been applied to predict P-E interactions [216]. However, CTCF was not a model feature in that study. In **Paper III**, we show that CTCF has in fact the highest predictive power for P-E interactions. This has also been corroborated by findings that CTCF (together with the cohesin subunit Rad21) contributes heavily to determine cell type-specific P-E interactions [222]. However, these authors do not differentiate between cell type-specific CTCF and constitutive CTCF binding. We used three instances of CTCF binding in our studies: CTCF counts at enhancer sites, cell type-specific CTCF counts and con-

stitutive CTCF counts between a P-E pair. We concluded that constitutive, rather than cell type-specific CTCF binding, is the strongest predictor of P-E interactions.

Recently, the role of CTCF binding motif orientation in the formation of TADs has been reported [143, 167, 245]. Interestingly, inversion of orientation of CTCF binding motifs using CRISPR/Cas9 genome editing leads to reconfiguration of the topological domain between an enhancer and its cognate promoter, resulting in a change in gene expression [246]. Adding direction of the CTCF binding motif as a feature for identification of P-E interactions would provide a stronger predicting power for P-E interactions than merely CTCF binding in our model. This remains to be tested.

Distance between an enhancer and its cognate promoter has previously been used to model P-E interactions [216]. It is also most common to use the closest promoter as the target of an enhancer [215]. However, there is a higher probability of finding a random interaction with shorter distance when using ChIA-PET or Hi-C data since chromosomes act as polymers [214]. Thus, inclusion of genomic distance can introduce a bias by learning subtle but artifactual differences in genomic distances between the true interactions and the computationally sampled non-interactions. Thus, we selected non-interacting P-E pairs with genomic distance as close as possible to the corresponding true interactions (**Paper III**) instead of adding genomic distance as a feature.

Sequence co-conservation has also been used as a feature to model P-E interactions [216]. The principle is that if an enhancer has to interact with a cognate promoter, there is an evolutionary incentive to prevent a loss of interaction capability. Studies suggest a higher probability for P-E interactions to be in a conserved synteny block [217, 218, 247]. Thus, rather than using genomic distance as a feature, positioning of a P-E pair in a synteny block could be used as a softer distance constraint.

TADs are conserved across cell-types [248] and could also be used as a feature in modeling P-E interactions. Indeed, interacting P-E pairs appear in the same TAD [249]. However, a reliance of TAD formation on CTCF-binding site direction has recently been shown [246]. Using CRISPR/Cas9-based genome editing in conjunction with 4C and Hi-C, a reversed orientation of CTCF-binding sites at protocadherin (*Pcdh*) and β-*globin* loci has been achieved, resulting in reorienta-

tion of TADs and in turn the loops between the enhancer and the gene loci; altering gene expression patterns [246]. Although this was not shown genome-wide, it appears that interaction of enhancers harboring CTCF sites with their cognate promoter can be determined by the underlying genomic sequence. Thus, using both CTCF binding direction and TADs may be redundant in the training model. On the other hand, TADs could be used as a soft distance parameter instead of or in combination with a conserved synteny block.

Gene expression levels have not been included as a feature of our RF model. This is a drawback since a P-E interaction increases promoter strength and cognate gene expression level. Since H3K27ac marks active enhancers, a correlation value between expression level and H3K27ac levels on the enhancer would provide a differentiating value between interacting and non-interacting P-E pairs. Thus, adding correlation between expression and enhancer H3K27ac levels would potentially increase the predictive power and accuracy of the model.

Enhancers were initially identified by finding regions of the DNA with a high density for TF binding sites [250]. Thus, ChIP-seq is increasingly being used to identify regions bound to TFs in order to locate context specific enhancers [118, 251,252]. P300 and CBP binding sites have been used to identify enhancers [253–255]. However, they may not be as informative as the combination of H3K4me1/2 and H3K27ac (used in **Paper III**) because P300 and CBP are found only on a subset of active enhancers [47, 49, 117, 122, 255]. We have therefore not used these features for enhancer identification (**Paper III**).

RNAPII at enhancers transcribes short enhancer RNA (eRNA) bi-directionally [256]; confirmed with bi-directional transcription from FANTOM5 TSS data [206, 215]. eRNAs have been used to measure activity of enhancers identified by H3K4me1 and showed positive correlation with expression levels of nearby genes [256]. eRNA maturation directly influences P-E interactions [257]. eRNA identification emerges as an accurate method of identifying active enhancers; however, the function of eRNA remains unexplored. It is also unknown if eRNAs are produced by all active enhancers, and eRNAs may also be cell type-specific, leaving quantification of P-E interactions using eRNAs context-dependent. Thus, even though H3K27ac does not always provide a strong correlation with enhancer activity [49, 117], it remains, together with H3K4me1/2, a robust alternative as a marker of enhancer activity. Prediction of enhancers and P-E interactions is

still in its infancy and will require more investigations to provide "true" positions of constitutive and cell type-specific enhancers. Predicted interacting P-E pairs identified in **Paper III** should therefore be considered as putative interactions.

4.3 Where does multivalency of chromatin states appear from?

We report that the enhancer mark H3K4me1 is retained during differentiation; however the H3K4me1/H3K27ac combination is more dynamic (**Paper I**), in line with studies showing cell type-specificity of H3K27ac [258, 259]. H3K4me1 marks enhancers which might be active or in a poised state whereas H3K27ac marks active enhancers. Thus during differentiation H3K27ac levels would be expected to vary based on the activity of the cognate gene [27]. We show that adipogenesis involves activation of genes in cohorts (**Paper I**); thus a reason to maintain H3K4me1 on enhancers may be to keep them "poised" to interact with a cognate promoter and modulate lineage-specific gene expression as differentiation proceeds.

The finding that the combination of H3K4me2 and H3K27me3 is the most dynamic during differentiation is striking (**Paper I**). This may be attributed to H3K27me3 per se: indeed we also identify an "H3K27me3-only" state as highly dynamic, while other states containing H3K4me2 are more stable. Adipogenic induction involves removal of H3K27me3 [260]. H3K4me2 and H3K27me3 co-incidence on adipogenic genes may represent an epigenetic "poising" for later transcriptional activation [260].

It should also be noted that our work relies on a cell population rather than single cells. Thus, it is unknown whether both marks occur on the same locus or whether they represent two distinct epigenetic marking (H3K4me2 or H3K27me3) in two cell sub-populations. Multivalency of histone PTMs can be tested by performing a sequential ChIP [260]; it would be informative to extend single-gene sequential ChIP analyses to a genome-wide mapping of co-incidence of H3K4me2 and H3K27me3 in an adipogenic context to provide information on multivalency.

4.4 The genic content of H2BGlcNAc domains (GADs): resolving the discrepancy

The difference reported in the genomic enrichment of H2BS112GlcNAc in **Paper II** and in a previous report [18] is in the genic content. We have shown that in ASCs and adipocytes, H2BGlcNAc localizes primarily in intergenic regions and that genes found in GADs are repressed (**Paper II**). In contrast, Fujiki et al. report, in HeLa cells, H2BGlcNAc mainly in genes that are transcriptionally active [18]. Several factors may explain this discrepancy.

One is the anti-H2BGlcNAc antibody used. Whereas Fujiki et al. [18] used a custom-made rabbit polyclonal antibody, we used the only H2BGlcNAc antibody available – a polyclonal antibody from Abcam raised against a 12 amino acid peptide of H2B GlcNAcylated on S112. We were not able to access the Fujiki antibody during our work. Thus, we have extensively characterized the Abcam antibody (**Paper II**) to conclude that the Abcam antibody is not entirely specific for S112GlcNAc as it appears to cross-react with an unmodified H2B peptide sequence. However, it is specific enough to ensure robustness of our ChIP data. It would be interesting to assess H2BGlcNAC profiles using the Fujiki antibody in ASCs, but this seems currently out of reach.

A second factor is the cell types examined. While Fujiki et al. [18] profiled H2BGlcNAc in HeLa cells (a polyploid cell line) we examined primary normal diploid cells [261] which could display a different GAD pattern. To address the issue of cell type-specificity of GADs, we have mapped by ChIP-seq using the Abcam antibody in HepG2 hepatoblasts lamin A/C LADs and H2BGlcNAc. To determine how dynamic lamin A/C LADs may be in relation to GADs, HepG2 cells were also treated with cyclosporin A (CspA) to induce an adipogenic phenotype. Using EDD, we show a redistribution of lamin A/C LADs after stimulation, and a change in the genomic properties of these LADs (namely, an increase in gene density; Figure 15A,B). Changes in LAD patterns are manifested by loss of LADs, gain of LADs or extension of existing LADs (Figure 15C). Thus, as in ASCs, lamin A/C LADs are remodeled during adipogenic stimulation of HepG2 cells; however in this system lamin A/C LADs tend to become surprisingly gene-rich after stimulation.

Next, a GAD overlap analysis (Figure 15A,D, column 2) indicates that GADs

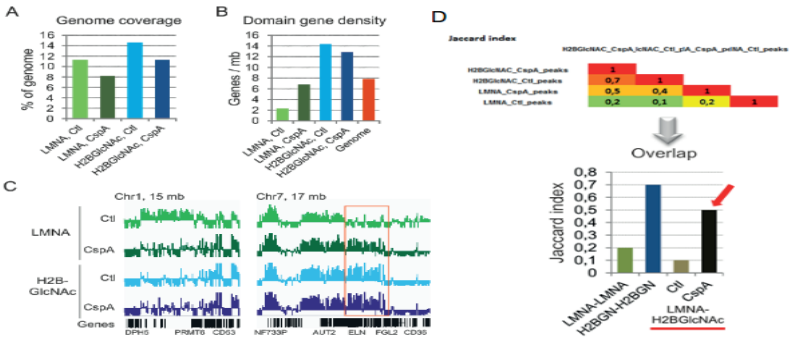

Figure 15: Characteristics of LADs and GADs in control and CspA-stimulated HepG2 cells. (A) Percent of the human genome covered by LADs and GADs. (B) Gene density in LADs and GADs. (C) IGV browser view of LMNA and H2BGlcNAc enrichment in a region of chromosomes 1 and 7 in control (Ctl) and CspA-stimulated HepG2 cells. Red frame highlights a region of de novo lamin A/C enrichment in CspA-treated cells, in an H2BGlcNAc-rich domain. (D) Jaccard index determination of the overlap between LADs and GADs in HepG2 cells before (Ctl) and after CspA treatment. Note the increase in lamin A/C (LMNA)-H2BGlcNAc overlap after CspA treatment (red arrow). A. Sørensen and P. Collas, unpublished.

are maintained between control and CspA-treated cells. Strikingly however, GAD gene density is higher from that of ASCs. Since we used the same antibody in both studies, this suggests that the genomic context of H2BGlcNAc varies between cell types and may represent a difference between primary cells and cell lines. The latter is supported by enrichment of H2BGlcNac in genes in HeLa cells [18]; however, more work is needed to validate this possibility, notably by ChIPing H2BGlcNAc from HeLa cells.

We further show in **Paper II** that during ASC differentiation, lamin A/C LADs form almost exclusively on GADs, suggesting that GADs pre-pattern *de novo* LAD formation. Our data from HepG2 cells show that a large proportion of *de novo* lamin A/C LADs form on GADs (Figure 15C,D). This suggests that as in ASCs, *de novo* lamin A/C LADs in HepG2 cells may be pre-patterned by GADs. We are currently testing whether this hypothesis is correct by altering GAD position and determining whether this affects lamin A/C LAD patterns.

4.5 Genome architecture, epigenetic marking and gene expression are regulators of cell fate: working models and future perspectives

In this thesis, we have used ASC differentiation from proliferation to adipocytes as a model to study changes in chromatin organization. Our main findings are highlighted in the model shown in Figure 16. We report that genes with stronger expression levels display greater variation of chromatin states. GeneA and GeneB have weak expression and thus, the loci are occupied by a limited number of histone PTM combinations (chromatin states; Figure 16A). After induction of differentiation, GeneA is strongly expressed, leading to high chromatin state dynamics of loci with many active chromatin states. In contrast, GeneB becomes repressed and shows fewer chromatin states; these include repressive modifications.

We report a method of predicting P-E interactions which reveals constitutive CTCF counts between an enhancer and promoter as a major factor for identification of P-E interactions. In Figure 16A, promoters of both GeneA and GeneB interact with their cognate enhancers. Active enhancers are marked by H3K4me1, H3K27ac and CTCF whereas inactive enhancers are only marked by H3K4me1. Additionally, CTCF counts between interacting P-E pairs are lower than non-interacting P-E pairs. However, after induction of differentiation, CTCF binding decreases between GeneA and the newly interacting enhancer. In contrast, CTCF binding increases between GeneB and the enhancer it lost contact with; resulting in GeneA gaining a P-E interaction whereas GeneB does not have any P-E interactions. Thus, one enhancer can regulate numerous promoters and at a given point in time a promoter can interact with multiple enhancers.

We further report that lamin A/C LADs are gene-rich in undifferentiated proliferating or cell cycle-arrested ASCs; LADs extend after cell cycle arrest and reform non-randomly on GADs after induction of differentiation; in contrast, GADs are stable during differentiation (Figure 16B). After adipogenic induction, *de novo* LADs become gene-poor and transcriptionally repressed. Moreover, as post-differentiation lamin A/C LADs overlap to greater extent with lamin B1 LADs [86], our results suggest the formation of these *de novo* lamin A/C LADs near the nuclear periphery. Our findings collectively provide new insights on how

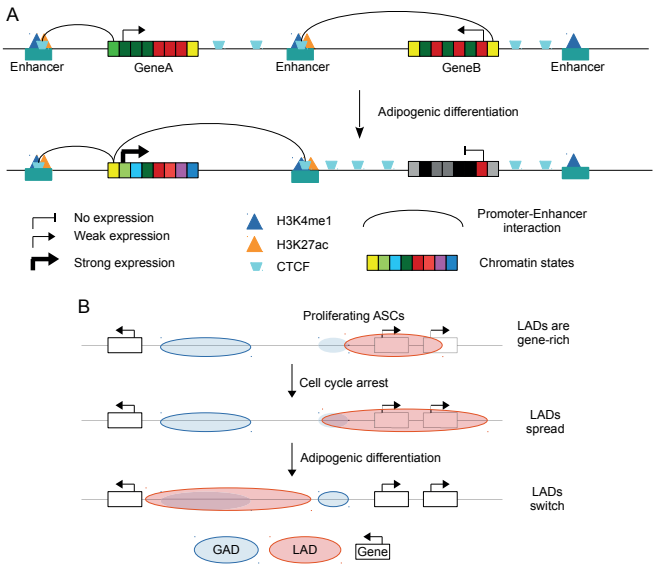

Figure 16: A model for changes in chromatin states and architecture during ASC differentiation. (A) Increased gene expression is coupled with varied chromatin states in a gene locus. Enhancers are marked by H3K4me1 while active enhancers harboring H3K27ac in addition. CTCF enrichment at the enhancer along with CTCF counts between a promoter and enhancer influence P-E interaction probability. (B) GADs are stable throughout differentiation. LADs are gene-rich in proliferating cells and expand with cell-cycle arrest. De novo LADs form non-randomly on GADs upon induction of adipogenesis.

histone PTMs, including domains of H2BS112GlcNAc (GADs), LADs and P-E interaction together regulate chromatin architecture and gene expression during stem cell differentiation.

In spite of our new findings, a number of issues are pending. (i) It remains unclear whether changes in chromatin states precede gene expression changes or whether gene expression levels impact local chromatin environment. (ii) Our data suggest that most of the genome harbors multiple coinciding histone PTMs; it will be important determine whether multivalency of histone marks does exist. (iii) We propose a model where GADs may pre-pattern differentiation-driven lamin A/C-chromatin associations. However, our data are at present correlative and one

will need to determine whether a functional relationship exists between LADs and GADs. This may be achieved by showing that removal or genomic displacement of H2BS112GlcNAc would prevent or alter *de novo* formation of LADs. A mutational approach can be envisaged, where a non-S112-GlcNAcylable H2B, such as an H2BS112A mutant, is expressed in order to replace endogenous H2B. This would enable assessing H2BS112GlcNAc levels and determining the impact on *de novo* lamin A/C LAD formation. Along these lines, it remains unknown whether the rearrangement of lamin A/C LADs is a prerequisite for ASC differentiation or a consequence thereof. (iv) We report that RF can be used to identify P-E interactions and that CTCF counts between promoters and enhancers are an important predictor of P-E interaction. How does our prediction relate to single nucleotide polymorphisms (SNPs) identified from genome-wide association studies in relation to genetic diseases? It is also unknown if P-E interactions occur due to epigenetic patterning or *vice-versa*.

Recent advances in molecular and cell biology techniques can be applied to advance our knowledge of epigenetic marking and genome organization. Fluorescence tagging of histone PTMs can be applied to track changes in histone PTMs *in vivo* [262]. In conjunction with nascent-transcript detection [263] or click chemistry [264], this can provide causal insights to the relationship between epigenetic marking and gene expression. m^6A-Tracer-derived techniques [84] and single-cell Hi-C-based modeling [168] at a high resolution provide information on locus positioning in the 3D nuclear space. These methods in conjunction with fluorescence *in situ* hybridization (FISH) or locus tracking in living cells using Cas9-EGFP tagging of loci [265] may shed light on the level of synchrony or stochasticity in chromatin organization as it relates to transcriptional status of individual loci in single cells. The parallel development of molecular techniques and of increasingly tailored and performant bioinformatic tools will lead to better understanding of chromatin regulation in dynamic systems.

Bibliography

[1] J. D. Watson, F. H. Crick, *et al.*, "Molecular structure of nucleic acids," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.

[2] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom, "A genomic code for nucleosome positioning," *Nature*, vol. 442, no. 7104, pp. 772–778, 2006.

[3] K. Van Holde, "Chromatin: Springer series in molecular biology," *New York, Berlin, Heidelberg London Paris Tokyo: Springer-Verlag*, 1988.

[4] A. Annunziato, "Dna packaging: nucleosomes and chromatin," *Nature Education*, vol. 1, no. 1, p. 26, 2008.

[5] D. E. Olins and A. L. Olins, "Chromatin history: our view from the bridge," *Nature reviews Molecular cell biology*, vol. 4, no. 10, pp. 809–814, 2003.

[6] T. G. Wolfsberg, J. McEntyre, and G. D. Schuler, "Guide to the draft human genome," *Nature*, vol. 409, no. 6822, pp. 824–826, 2001.

[7] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigó, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, *et al.*, "Identification and analysis of functional elements in 1% of the human genome by the encode pilot project," *Nature*, vol. 447, no. 7146, pp. 799–816, 2007.

[8] J. van Arensbergen, B. van Steensel, and H. J. Bussemaker, "In search of the determinants of enhancer–promoter interaction specificity," *Trends in cell biology*, vol. 24, no. 11, pp. 695–702, 2014.

[9] M. S. Kowalczyk, J. R. Hughes, D. Garrick, M. D. Lynch, J. A. Sharpe, J. A. Sloane-Stanley, S. J. McGowan, M. De Gobbi, M. Hosseini, D. Vernimmen, *et al.*, "Intragenic enhancers act as alternative promoters," *Molecular cell*, vol. 45, no. 4, pp. 447–458, 2012.

[10] A. D. Goldberg, C. D. Allis, and E. Bernstein, "Epigenetics: a landscape takes shape," *Cell*, vol. 128, no. 4, pp. 635–638, 2007.

[11] C. H. Waddington *et al.*, "The strategy of the genes. a discussion of some aspects of theoretical biology. with an appendix by h. kacser.," *The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser.*, pp. ix+–262, 1957.

[12] M. Grunstein, "Histone acetylation in chromatin structure and transcription," *Nature*, vol. 389, no. 6649, pp. 349–352, 1997.

[13] G. L. Cuthbert, S. Daujat, A. W. Snowden, H. Erdjument-Bromage, T. Hagiwara, M. Yamada, R. Schneider, P. D. Gregory, P. Tempst, A. J. Bannister, *et al.*, "Histone deimination antagonizes arginine methylation," *Cell*, vol. 118, no. 5, pp. 545–553, 2004.

[14] J. C. Rice and C. D. Allis, "Histone methylation versus histone acetylation: new insights into epigenetic regulation," *Current opinion in cell biology*, vol. 13, no. 3, pp. 263–273, 2001.

[15] W.-S. Lo, R. C. Trievel, J. R. Rojas, L. Duggan, J.-Y. Hsu, C. D. Allis, R. Marmorstein, and S. L. Berger, "Phosphorylation of serine 10 in histone h3 is functionally linked in vitro and in vivo to gcn5-mediated acetylation at lysine 14," *Molecular cell*, vol. 5, no. 6, pp. 917–926, 2000.

[16] A. E. Kelly, C. Ghenoiu, J. Z. Xue, C. Zierhut, H. Kimura, and H. Funabiki, "Survivin reads phosphorylated histone h3 threonine 3 to activate the mitotic kinase aurora b," *Science*, vol. 330, no. 6001, pp. 235–239, 2010.

[17] D. Nathan, K. Ingvarsdottir, D. E. Sterner, G. R. Bylebyl, M. Dokmanovic, J. A. Dorsey, K. A. Whelan, M. Krsmanovic, W. S. Lane, P. B. Meluh, *et al.*, "Histone sumoylation is a negative regulator in saccharomyces cerevisiae

and shows dynamic interplay with positive-acting histone modifications," *Genes & development*, vol. 20, no. 8, pp. 966–976, 2006.

[18] R. Fujiki, W. Hashiba, H. Sekine, A. Yokoyama, T. Chikanishi, S. Ito, Y. Imai, J. Kim, H. H. He, K. Igarashi, *et al.*, "Glcnacylation of histone h2b facilitates its monoubiquitination," *Nature*, vol. 480, no. 7378, pp. 557–560, 2011.

[19] T. Kouzarides, "Chromatin modifications and their function," *Cell*, vol. 128, no. 4, pp. 693–705, 2007.

[20] H. Santos-Rosa, A. Kirmizis, C. Nelson, T. Bartke, N. Saksouk, J. Cote, and T. Kouzarides, "Histone h3 tail clipping regulates gene expression," *Nature structural & molecular biology*, vol. 16, no. 1, pp. 17–22, 2009.

[21] V. W. Zhou, A. Goren, and B. E. Bernstein, "Charting histone modifications and the functional organization of mammalian genomes," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 7–18, 2011.

[22] S. Lall, "Primers on chromatin," *Nature structural & molecular biology*, vol. 14, no. 11, pp. 1110–1115, 2007.

[23] B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, *et al.*, "A bivalent chromatin structure marks key developmental genes in embryonic stem cells," *Cell*, vol. 125, no. 2, pp. 315–326, 2006.

[24] S. Feng, S. J. Cokus, X. Zhang, P.-Y. Chen, M. Bostick, M. G. Goll, J. Hetzel, J. Jain, S. H. Strauss, M. E. Halpern, *et al.*, "Conservation and divergence of methylation patterning in plants and animals," *Proceedings of the National Academy of Sciences*, vol. 107, no. 19, pp. 8689–8694, 2010.

[25] C. Y. Okitsu, J. C. F. Hsieh, and C.-L. Hsieh, "Transcriptional activity affects the h3k4me3 level and distribution in the coding region," *Molecular and cellular biology*, vol. 30, no. 12, pp. 2933–2946, 2010.

[26] T. S. Mikkelsen, Z. Xu, X. Zhang, L. Wang, J. M. Gimble, E. S. Lander, and E. D. Rosen, "Comparative epigenomic analysis of murine and human adipogenesis," *Cell*, vol. 143, no. 1, pp. 156–169, 2010.

[27] M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, *et al.*, "Histone h3k27ac separates active from poised enhancers and predicts developmental state," *Proceedings of the National Academy of Sciences*, vol. 107, no. 50, pp. 21931–21936, 2010.

[28] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao, "High-resolution profiling of histone methylations in the human genome," *Cell*, vol. 129, no. 4, pp. 823–837, 2007.

[29] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.-K. Kim, R. P. Koche, *et al.*, "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells," *Nature*, vol. 448, no. 7153, pp. 553–560, 2007.

[30] H. H. Ng, F. Robert, R. A. Young, and K. Struhl, "Targeted recruitment of set1 histone methylase by elongating pol ii provides a localized mark and memory of recent transcriptional activity," *Molecular cell*, vol. 11, no. 3, pp. 709–719, 2003.

[31] J. R. Tollervey and V. V. Lunyak, "Epigenetics: judge, jury and executioner of stem cell fate," *Epigenetics*, vol. 7, no. 8, pp. 823–840, 2012.

[32] T. Cremer and C. Cremer, "Chromosome territories, nuclear architecture and gene regulation in mammalian cells," *Nature reviews genetics*, vol. 2, no. 4, pp. 292–301, 2001.

[33] N. Gilbert, S. Boyle, H. Fiegler, K. Woodfine, N. P. Carter, and W. A. Bickmore, "Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers," *Cell*, vol. 118, no. 5, pp. 555–566, 2004.

[34] R. P. McCord, A. Nazario-Toole, H. Zhang, P. S. Chines, Y. Zhan, M. R. Erdos, F. S. Collins, J. Dekker, and K. Cao, "Correlated alterations in genome organization, histone methylation, and dna–lamin a/c interactions in hutchinson-gilford progeria syndrome," *Genome research*, vol. 23, no. 2, pp. 260–269, 2013.

[35] M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, *et al.*, "Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals," *Nature*, vol. 458, no. 7235, pp. 223–227, 2009.

[36] T. R. Mercer, M. E. Dinger, and J. S. Mattick, "Long non-coding rnas: insights into functions," *Nature Reviews Genetics*, vol. 10, no. 3, pp. 155–159, 2009.

[37] M. Esteller, "Non-coding rnas in human disease," *Nature Reviews Genetics*, vol. 12, no. 12, pp. 861–874, 2011.

[38] A. Postepska-Igielska, D. Krunic, N. Schmitt, K. M. Greulich-Bode, P. Boukamp, and I. Grummt, "The chromatin remodelling complex norc safeguards genome stability by heterochromatin formation at telomeres and centromeres," *EMBO reports*, vol. 14, no. 8, pp. 704–710, 2013.

[39] K. S. Bloom, "Centromeric heterochromatin: the primordial segregation machine.," *Annual review of genetics*, vol. 48, pp. 457–484, 2013.

[40] K.-i. Noma, C. D. Allis, and S. I. Grewal, "Transitions in distinct histone h3 methylation patterns at the heterochromatin domain boundaries," *Science*, vol. 293, no. 5532, pp. 1150–1155, 2001.

[41] J. Ren and R. A. Martienssen, "Silent decision: Hp1 protein escorts heterochromatic rnas to their destiny," *The EMBO journal*, vol. 31, no. 15, pp. 3237–3238, 2012.

[42] A. B. Brinkman, T. Roelofsen, S. W. Pennings, J. H. Martens, T. Jenuwein, and H. G. Stunnenberg, "Histone modification patterns associated with the human x chromosome," *EMBO reports*, vol. 7, no. 6, pp. 628–634, 2006.

[43] I. Solovei, A. S. Wang, K. Thanisch, C. S. Schmidt, S. Krebs, M. Zwerger, T. V. Cohen, D. Devys, R. Foisner, L. Peichl, *et al.*, "Lbr and lamin a/c sequentially tether peripheral heterochromatin and inversely regulate differentiation," *Cell*, vol. 152, no. 3, pp. 584–598, 2013.

[44] K. I. Lio, A. Clarke, and K. Reed, "Characterizing the role of heterochromatin protein 1 gamma in normal intestinal homeostasis and tumorigenesis," *Cancer Research*, vol. 74, no. 19 Supplement, pp. 381–381, 2014.

[45] N. Uranova, D. Orlovskaya, O. Vikhreva, I. Zimina, N. Kolomeets, V. Vostrikov, and V. Rachmanova, "Electron microscopy of oligodendroglia in severe mental illness," *Brain research bulletin*, vol. 55, no. 5, pp. 597–610, 2001.

[46] G. J. Filion, J. G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L. D. Ward, W. Brugman, I. J. de Castro, R. M. Kerkhoven, H. J. Bussemaker, *et al.*, "Systematic protein location mapping reveals five principal chromatin types in drosophila cells," *Cell*, vol. 143, no. 2, pp. 212–224, 2010.

[47] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, *et al.*, "Mapping and analysis of chromatin state dynamics in nine human cell types," *Nature*, vol. 473, no. 7345, pp. 43–49, 2011.

[48] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble, "Unsupervised pattern discovery in human chromatin structure through genomic segmentation," *Nature methods*, vol. 9, no. 5, pp. 473–476, 2012.

[49] S. Bonn, R. P. Zinzen, C. Girardot, E. H. Gustafson, A. Perez-Gonzalez, N. Delhomme, Y. Ghavi-Helm, B. Wilczyński, A. Riddell, and E. E. Furlong, "Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development," *Nature genetics*, vol. 44, no. 2, pp. 148–156, 2012.

[50] S. T. Keating and A. El-Osta, "Epigenetics and metabolism," *Circulation research*, vol. 116, no. 4, pp. 715–736, 2015.

[51] J. A. Hanover, M. W. Krause, and D. C. Love, "Bittersweet memories: linking metabolism to epigenetics through o-glcnacylation," *Nature Reviews Molecular Cell Biology*, vol. 13, no. 5, pp. 312–321, 2012.

[52] S. Marshall, W. Garvey, and R. Traxinger, "New insights into the metabolic regulation of insulin action and insulin resistance: role of glucose and amino acids.," *The FASEB Journal*, vol. 5, no. 15, pp. 3031–3036, 1991.

[53] L. K. Kreppel, M. A. Blomberg, and G. W. Hart, "Dynamic glycosylation of nuclear and cytosolic proteins cloning and characterization of a unique o-glcnac transferase with multiple tetratricopeptide repeats," *Journal of Biological Chemistry*, vol. 272, no. 14, pp. 9308–9315, 1997.

[54] Y. Gao, L. Wells, F. I. Comer, G. J. Parker, and G. W. Hart, "Dynamic o-glycosylation of nuclear and cytosolic proteins cloning and characterization of a neutral, cytosolic β-n-acetylglucosaminidase from human brain," *Journal of Biological Chemistry*, vol. 276, no. 13, pp. 9838–9845, 2001.

[55] M. C. Gambetta and J. Müller, "A critical perspective of the diverse roles of o-glcnac transferase in chromatin," *Chromosoma*, pp. 1–14, 2015.

[56] J. J. Fong, B. L. Nguyen, R. Bridger, E. E. Medrano, L. Wells, S. Pan, and R. N. Sifers, "β-n-acetylglucosamine (o-glcnac) is a novel regulator of mitosis-specific phosphorylations on histone h3," *Journal of Biological Chemistry*, vol. 287, no. 15, pp. 12195–12203, 2012.

[57] P. Vella, A. Scelfo, S. Jammula, F. Chiacchiera, K. Williams, A. Cuomo, A. Roberto, J. Christensen, T. Bonaldi, K. Helin, *et al.*, "Tet proteins connect the o-linked n-acetylglucosamine transferase ogt to chromatin in embryonic stem cells," *Molecular cell*, vol. 49, no. 4, pp. 645–656, 2013.

[58] S. Zhang, K. Roche, H.-P. Nasheuer, and N. F. Lowndes, "Modification of histones by sugar β-n-acetylglucosamine (glcnac) occurs on multiple residues, including histone h3 serine 10, and is cell cycle-regulated," *Journal of Biological Chemistry*, vol. 286, no. 43, pp. 37483–37495, 2011.

[59] Q. Chen, Y. Chen, C. Bian, R. Fujiki, and X. Yu, "Tet2 promotes histone o-glcnacylation during gene transcription," *Nature*, vol. 493, no. 7433, pp. 561–564, 2013.

[60] R. Deplus, B. Delatte, M. K. Schwinn, M. Defrance, J. Méndez, N. Murphy, M. A. Dawson, M. Volkmar, P. Putmans, E. Calonne, *et al.*, "Tet2

and tet3 regulate glcnacylation and h3k4 methylation through ogt and set1/compass," *The EMBO journal*, vol. 32, no. 5, pp. 645–655, 2013.

[61] M. C. Gambetta, K. Oktaba, and J. Müller, "Essential role of the glycosyltransferase sxc/ogt in polycomb repression," *Science*, vol. 325, no. 5936, pp. 93–96, 2009.

[62] M. C. Gambetta and J. Müller, "O-glcnacylation prevents aggregation of the polycomb group repressor polyhomeotic," *Developmental cell*, vol. 31, no. 5, pp. 629–639, 2014.

[63] S. Özcan, S. S. Andrali, and J. E. Cantrell, "Modulation of transcription factor function by o-glcnac modification," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, vol. 1799, no. 5, pp. 353–364, 2010.

[64] C.-S. Chu, P.-W. Lo, Y.-H. Yeh, P.-H. Hsu, S.-H. Peng, Y.-C. Teng, M.-L. Kang, C.-H. Wong, and L.-J. Juan, "O-glcnacylation regulates ezh2 protein stability and function," *Proceedings of the National Academy of Sciences*, vol. 111, no. 4, pp. 1355–1360, 2014.

[65] X. Yang, F. Zhang, and J. E. Kudlow, "Recruitment of o-glcnac transferase to promoters by corepressor msin3a: coupling protein o-glcnacylation to transcriptional repression," *Cell*, vol. 110, no. 1, pp. 69–80, 2002.

[66] V. Butin-Israeli, S. A. Adam, A. E. Goldman, and R. D. Goldman, "Nuclear lamin functions and disease," *Trends in genetics*, vol. 28, no. 9, pp. 464–471, 2012.

[67] L. Gerace and M. D. Huber, "Nuclear lamina at the crossroads of the cytoplasm and nucleus," *Journal of structural biology*, vol. 177, no. 1, pp. 24–31, 2012.

[68] B. Burke and C. L. Stewart, "The nuclear lamins: flexibility in function," *Nature reviews Molecular cell biology*, vol. 14, no. 1, pp. 13–24, 2013.

[69] C. Stewart and B. Burke, "Teratocarcinoma stem cells and early mouse embryos contain only a single major lamin polypeptide closely resembling lamin b," *Cell*, vol. 51, no. 3, pp. 383–392, 1987.

[70] H. J. Worman, I. Lazaridis, and S. Georgatos, "Nuclear lamina heterogeneity in mammalian cells. differential expression of the major lamins and variations in lamin b phosphorylation.," *Journal of Biological Chemistry*, vol. 263, no. 24, pp. 12135–12141, 1988.

[71] E. Delbarre, M. Tramier, M. Coppey-Moisan, C. Gaillard, J.-C. Courvalin, and B. Buendia, "The truncated prelamin a in hutchinson–gilford progeria syndrome alters segregation of a-type and b-type lamin homopolymers," *Human molecular genetics*, vol. 15, no. 7, pp. 1113–1122, 2006.

[72] T. Kolb, K. Maaß, M. Hergt, U. Aebi, and H. Herrmann, "Lamin a and lamin c form homodimers and coexist in higher complex forms both in the nucleoplasmic fraction and in the lamina of cultured human cells," *Nucleus*, vol. 2, no. 5, pp. 425–433, 2011.

[73] T. Dechat, K. Gesson, and R. Foisner, "Lamina-independent lamins in the nuclear interior serve important functions," in *Cold Spring Harbor symposia on quantitative biology*, vol. 75, pp. 533–543, Cold Spring Harbor Laboratory Press, 2010.

[74] T. Dechat, B. Korbei, O. A. Vaughan, S. Vlcek, C. J. Hutchison, and R. Foisner, "Lamina-associated polypeptide 2alpha binds intranuclear a-type lamins," *Journal of Cell Science*, vol. 113, no. 19, pp. 3473–3484, 2000.

[75] N. Naetar and R. Foisner, "Lamin complexes in the nuclear interior control progenitor cell proliferation and tissue homeostasis," *Cell Cycle*, vol. 8, no. 10, pp. 1488–1493, 2009.

[76] B. R. Johnson, R. T. Nitta, R. L. Frock, L. Mounkes, D. A. Barbie, C. L. Stewart, E. Harlow, and B. K. Kennedy, "A-type lamins regulate retinoblastoma protein function by promoting subnuclear localization and preventing proteasomal degradation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 26, pp. 9677–9682, 2004.

[77] N. Naetar, B. Korbei, S. Kozlov, M. A. Kerenyi, D. Dorner, R. Kral, I. Gotic, P. Fuchs, T. V. Cohen, R. Bittner, *et al.*, "Loss of nucleoplasmic lap2α–lamin a complexes causes erythroid and epidermal progenitor hyperproliferation," *nature cell biology*, vol. 10, no. 11, pp. 1341–1348, 2008.

[78] I. Gotic and R. Foisner, "Multiple novel functions of lamina associated polypeptide 2α in striated muscle," *Nucleus*, vol. 1, no. 5, pp. 397–401, 2010.

[79] D. Dorner, S. Vlcek, N. Foeger, A. Gajewski, C. Makolm, J. Gotzmann, C. J. Hutchison, and R. Foisner, "Lamina-associated polypeptide 2α regulates cell cycle progression and differentiation via the retinoblastoma–e2f pathway," *The Journal of cell biology*, vol. 173, no. 1, pp. 83–93, 2006.

[80] K. Gesson, S. Vidak, and R. Foisner, "Lamina-associated polypeptide (lap) 2α and nucleoplasmic lamins in adult stem cell regulation and disease," in *Seminars in cell & developmental biology*, vol. 29, pp. 116–124, Elsevier, 2014.

[81] N. Kubben, M. Adriaens, W. Meuleman, J. W. Voncken, B. van Steensel, and T. Misteli, "Mapping of lamin a-and progerin-interacting genome regions," *Chromosoma*, vol. 121, no. 5, pp. 447–464, 2012.

[82] E. Lund, A. R. Oldenburg, E. Delbarre, C. T. Freberg, I. Duband-Goulet, R. Eskeland, B. Buendia, and P. Collas, "Lamin a/c-promoter interactions specify chromatin state–dependent transcription outcomes," *Genome research*, vol. 23, no. 10, pp. 1580–1589, 2013.

[83] V. Stierlé, J. Couprie, C. Östlund, I. Krimm, S. Zinn-Justin, P. Hossenlopp, H. J. Worman, J.-C. Courvalin, and I. Duband-Goulet, "The carboxyl-terminal region common to lamins a and c contains a dna binding domain," *Biochemistry*, vol. 42, no. 17, pp. 4819–4828, 2003.

[84] J. Kind and B. van Steensel, "Genome–nuclear lamina interactions and gene regulation," *Current opinion in cell biology*, vol. 22, no. 3, pp. 320–325, 2010.

[85] H. Pickersgill, B. Kalverda, E. de Wit, W. Talhout, M. Fornerod, and B. van Steensel, "Characterization of the drosophila melanogaster genome at the nuclear lamina," *Nature genetics*, vol. 38, no. 9, pp. 1005–1014, 2006.

[86] L. Guelen, L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat, *et al.*, "Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions," *Nature*, vol. 453, no. 7197, pp. 948–951, 2008.

[87] D. Peric-Hupkes, W. Meuleman, L. Pagie, S. W. Bruggeman, I. Solovei, W. Brugman, S. Gräf, P. Flicek, R. M. Kerkhoven, M. van Lohuizen, *et al.*, "Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation," *Molecular cell*, vol. 38, no. 4, pp. 603–613, 2010.

[88] J. G. van Bemmel, L. Pagie, U. Braunschweig, W. Brugman, W. Meuleman, R. M. Kerkhoven, and B. van Steensel, "The insulator protein su (hw) fine-tunes nuclear lamina interactions of the drosophila genome," *PLoS One*, vol. 5, no. 11, p. e15013, 2010.

[89] J. Kind, L. Pagie, H. Ortabozkoyun, S. Boyle, S. S. de Vries, H. Janssen, M. Amendola, L. D. Nolen, W. A. Bickmore, and B. van Steensel, "Single-cell dynamics of genome-nuclear lamina interactions," *Cell*, vol. 153, no. 1, pp. 178–192, 2013.

[90] M. Sadaie, R. Salama, T. Carroll, K. Tomimatsu, T. Chandra, A. R. Young, M. Narita, P. A. Pérez-Mancera, D. C. Bennett, H. Chong, *et al.*, "Redistribution of the lamin b1 genomic binding profile affects rearrangement of heterochromatic domains and sahf formation during senescence," *Genes & development*, vol. 27, no. 16, pp. 1800–1808, 2013.

[91] P. P. Shah, G. Donahue, G. L. Otte, B. C. Capell, D. M. Nelson, K. Cao, V. Aggarwala, H. A. Cruickshanks, T. S. Rai, T. McBryan, *et al.*, "Lamin b1 depletion in senescent cells triggers large-scale changes in gene expression and the chromatin landscape," *Genes & development*, vol. 27, no. 16, pp. 1787–1799, 2013.

[92] W. Meuleman, D. Peric-Hupkes, J. Kind, J.-B. Beaudry, L. Pagie, M. Kellis, M. Reinders, L. Wessels, and B. van Steensel, "Constitutive nuclear lamina–genome interactions are highly conserved and associated with a/t-rich sequence," *Genome research*, vol. 23, no. 2, pp. 270–280, 2013.

[93] E. Lund, A. R. Oldenburg, and P. Collas, "Enriched domain detector: a program for detection of wide genomic enrichment domains robust against local variations," *Nucleic acids research*, vol. 42, no. 11, pp. e92–e92, 2014.

[94] K. Reddy, J. Zullo, E. Bertolino, and H. Singh, "Transcriptional repression mediated by repositioning of genes to the nuclear lamina," *Nature*, vol. 452, no. 7184, pp. 243–247, 2008.

[95] A. Mattout, B. L. Pike, B. D. Towbin, E. M. Bank, A. Gonzalez-Sandoval, M. B. Stadler, P. Meister, Y. Gruenbaum, and S. M. Gasser, "An edmd mutation in c. elegans lamin blocks muscle-specific gene relocation and compromises muscle integrity," *Current Biology*, vol. 21, no. 19, pp. 1603–1614, 2011.

[96] B. D. Towbin, C. González-Aguilera, R. Sack, D. Gaidatzis, V. Kalck, P. Meister, P. Askjaer, and S. M. Gasser, "Step-wise methylation of histone h3k9 positions heterochromatin at the nuclear periphery," *Cell*, vol. 150, no. 5, pp. 934–947, 2012.

[97] J. M. Zullo, I. A. Demarco, R. Piqué-Regi, D. J. Gaffney, C. B. Epstein, C. J. Spooner, T. R. Luperchio, B. E. Bernstein, J. K. Pritchard, K. L. Reddy, *et al.*, "Dna sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina," *Cell*, vol. 149, no. 7, pp. 1474–1487, 2012.

[98] J. C. Harr, T. R. Luperchio, X. Wong, E. Cohen, S. J. Wheelan, and K. L. Reddy, "Directed targeting of chromatin to the nuclear lamina is mediated by chromatin state and a-type lamins," *The Journal of cell biology*, vol. 208, no. 1, pp. 33–52, 2015.

[99] N. Zuleger, S. Boyle, D. A. Kelly, J. I. de Las Heras, V. Lazou, N. Korfali, D. G. Batrakou, K. N. Randles, G. E. Morris, D. J. Harrison, *et al.*, "Specific nuclear envelope transmembrane proteins can promote the location of chromosomes to and from the nuclear periphery," *Genome biology*, vol. 14, no. 2, p. R14, 2013.

[100] N. Zuleger, M. I. Robson, and E. C. Schirmer, "The nuclear envelope as a chromatin organizer," *Nucleus*, vol. 2, no. 5, pp. 339–349, 2011.

[101] V. Andrés and J. M. González, "Role of a-type lamins in signaling, transcription, and chromatin organization," *The Journal of cell biology*, vol. 187, no. 7, pp. 945–957, 2009.

[102] K. H. Schreiber and B. K. Kennedy, "When lamins go bad: nuclear structure and disease," *Cell*, vol. 152, no. 6, pp. 1365–1375, 2013.

[103] G. G. Gundersen and H. J. Worman, "Nuclear positioning," *Cell*, vol. 152, no. 6, pp. 1376–1389, 2013.

[104] H. J. Worman, "Nuclear lamins and laminopathies," *The Journal of pathology*, vol. 226, no. 2, pp. 316–325, 2012.

[105] C. Y. Ho, D. E. Jaalouk, and J. Lammerding, "Novel insights into the disease etiology of laminopathies," *Rare Diseases*, vol. 1, no. 1, pp. 507–11, 2013.

[106] J. C. Choi and H. J. Worman, "Reactivation of autophagy ameliorates lmna cardiomyopathy," *Autophagy*, vol. 9, no. 1, pp. 110–111, 2013.

[107] D. M. Lehman, D.-J. Fu, A. B. Freeman, K. J. Hunt, R. J. Leach, T. Johnson-Pais, J. Hamlington, T. D. Dyer, R. Arya, H. Abboud, *et al.*, "A single nucleotide polymorphism in mgea5 encoding o-glcnac–selective n-acetyl-β-d glucosaminidase is associated with type 2 diabetes in mexican americans," *Diabetes*, vol. 54, no. 4, pp. 1214–1221, 2005.

[108] E. Smith and A. Shilatifard, "Enhancer biology and enhanceropathies," *Nature structural & molecular biology*, vol. 21, no. 3, pp. 210–219, 2014.

[109] Y. Zhang, C.-H. Wong, R. Y. Birnbaum, G. Li, R. Favaro, C. Y. Ngan, J. Lim, E. Tai, H. M. Poh, E. Wong, *et al.*, "Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations," *Nature*, vol. 504, no. 7479, pp. 306–310, 2013.

[110] S. F. Schmidt, B. D. Larsen, A. Loft, R. Nielsen, J. G. S. Madsen, and S. Mandrup, "Acute tnf-induced repression of cell identity genes is mediated by nfκb-directed redistribution of cofactors from super-enhancers," *Genome research*, vol. 25, no. 9, pp. 1281–1294, 2015.

[111] F. Spitz and E. E. Furlong, "Transcription factors: from enhancer binding to developmental control," *Nature Reviews Genetics*, vol. 13, no. 9, pp. 613–626, 2012.

[112] K. M. Lower, J. R. Hughes, M. De Gobbi, S. Henderson, V. Viprakasit, C. Fisher, A. Goriely, H. Ayyub, J. Sloane-Stanley, D. Vernimmen, *et al.*, "Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition," *Proceedings of the National Academy of Sciences*, vol. 106, no. 51, pp. 21771–21776, 2009.

[113] A. Sanyal, B. R. Lajoie, G. Jain, and J. Dekker, "The long-range interaction landscape of gene promoters," *Nature*, vol. 489, no. 7414, pp. 109–113, 2012.

[114] M. Agelopoulos, D. J. McKay, and R. S. Mann, "Developmental regulation of chromatin conformation by hox proteins in drosophila," *Cell reports*, vol. 1, no. 4, pp. 350–359, 2012.

[115] G. Andrey, T. Montavon, B. Mascrez, F. Gonzalez, D. Noordermeer, M. Leleu, D. Trono, F. Spitz, and D. Duboule, "A switch between topological domains underlies hoxd genes collinearity in mouse limbs," *Science*, vol. 340, no. 6137, p. 1234167, 2013.

[116] F. Jin, Y. Li, J. R. Dixon, S. Selvaraj, Z. Ye, A. Y. Lee, C.-A. Yen, A. D. Schmitt, C. A. Espinoza, and B. Ren, "A high-resolution map of the three-dimensional chromatin interactome in human cells," *Nature*, vol. 503, no. 7475, pp. 290–294, 2013.

[117] C. D. Arnold, D. Gerlach, C. Stelzer, Ł. M. Boryń, M. Rath, and A. Stark, "Genome-wide quantitative enhancer activity maps identified by starr-seq," *Science*, vol. 339, no. 6123, pp. 1074–1077, 2013.

[118] D. Shlyueva, C. Stelzer, D. Gerlach, J. O. Yáñez-Cuna, M. Rath, Ł. M. Boryń, C. D. Arnold, and A. Stark, "Hormone-responsive enhancer-activity maps reveal predictive motifs, indirect repression, and targeting of closed chromatin," *Molecular cell*, vol. 54, no. 1, pp. 180–192, 2014.

[119] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, no. 7398, pp. 376–380, 2012.

[120] G. Li, M. Fullwood, H. Xu, F. Mulawadi, S. Velkov, V. Vega, P. Ariyaratne, Y. Mohamed, H. Ooi, C. Tennakoon, *et al.*, "Software chia-pet tool for comprehensive chromatin interaction analysis with paired-end tag sequencing," *Genome Biol*, vol. 11, p. R22, 2010.

[121] D. J. Fitzpatrick, C. J. Ryan, N. Shah, D. Greene, C. Molony, and D. C. Shields, "Genome-wide epistatic expression quantitative trait loci discovery in four human tissues reveals the importance of local chromosomal interactions governing gene expression," *BMC genomics*, vol. 16, no. 1, p. 109, 2015.

[122] A. Rada-Iglesias, R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn, and J. Wysocka, "A unique chromatin signature uncovers early developmental enhancers in humans," *Nature*, vol. 470, no. 7333, pp. 279–283, 2011.

[123] G. E. Zentner, P. J. Tesar, and P. C. Scacheri, "Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions," *Genome research*, vol. 21, no. 8, pp. 1273–1283, 2011.

[124] K. S. Sandhu, G. Li, H. M. Poh, Y. L. K. Quek, Y. Y. Sia, S. Q. Peh, F. H. Mulawadi, J. Lim, M. Sikic, F. Menghi, *et al.*, "Large-scale functional organization of long-range chromatin interaction networks," *Cell reports*, vol. 2, no. 5, pp. 1207–1219, 2012.

[125] J. Zhang, H. M. Poh, S. Q. Peh, Y. Y. Sia, G. Li, F. H. Mulawadi, Y. Goh, M. J. Fullwood, W.-K. Sung, X. Ruan, *et al.*, "Chia-pet analysis of transcriptional chromatin interactions," *Methods*, vol. 58, no. 3, pp. 289–299, 2012.

[126] K.-R. Kieffer-Kwon, Z. Tang, E. Mathe, J. Qian, M.-H. Sung, G. Li, W. Resch, S. Baek, N. Pruett, L. Grøntved, *et al.*, "Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation," *Cell*, vol. 155, no. 7, pp. 1507–1520, 2013.

[127] R. Kellum and P. Schedl, "A position-effect assay for boundaries of higher order chromosomal domains," *Cell*, vol. 64, no. 5, pp. 941–950, 1991.

[128] A. M. Wood, K. Van Bortle, E. Ramos, N. Takenaka, M. Rohrbaugh, B. C. Jones, K. C. Jones, and V. G. Corces, "Regulation of chromatin organization and inducible gene expression by a drosophila insulator," *Molecular cell*, vol. 44, no. 1, pp. 29–38, 2011.

[129] R. Ohlsson, R. Renkawitz, and V. Lobanenkov, "Ctcf is a uniquely versatile transcription regulator linked to epigenetics and disease," *TRENDS in Genetics*, vol. 17, no. 9, pp. 520–527, 2001.

[130] H. Chen, Y. Tian, W. Shu, X. Bo, and S. Wang, "Comprehensive identification and annotation of cell type-specific and ubiquitous ctcf-binding sites in the human genome," *PloS one*, vol. 7, no. 7, p. e41374, 2012.

[131] E. P. Consortium *et al.*, "An integrated encyclopedia of dna elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[132] S. Cuddapah, R. Jothi, D. E. Schones, T.-Y. Roh, K. Cui, and K. Zhao, "Global analysis of the insulator binding protein ctcf in chromatin barrier regions reveals demarcation of active and repressive domains," *Genome research*, vol. 19, no. 1, pp. 24–32, 2009.

[133] Z. Zhao, G. Tavoosidana, M. Sjölinder, A. Göndör, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. Sandhu, U. Singh, *et al.*, "Circular chromosome conformation capture (4c) uncovers extensive networks of epigenet-

ically regulated intra-and interchromosomal interactions," *Nature genetics*, vol. 38, no. 11, pp. 1341–1347, 2006.

[134] E. Splinter, H. Heath, J. Kooren, R.-J. Palstra, P. Klous, F. Grosveld, N. Galjart, and W. de Laat, "Ctcf mediates long-range chromatin looping and local histone modification in the β-globin locus," *Genes & development*, vol. 20, no. 17, pp. 2349–2354, 2006.

[135] C. A. Espinoza and B. Ren, "Mapping higher order structure of chromatin domains," *Nature genetics*, vol. 43, no. 7, pp. 615–616, 2011.

[136] L. Handoko, H. Xu, G. Li, C. Y. Ngan, E. Chew, M. Schnapp, C. W. H. Lee, C. Ye, J. L. H. Ping, F. Mulawadi, *et al.*, "Ctcf-mediated functional chromatin interactome in pluripotent cells," *Nature genetics*, vol. 43, no. 7, pp. 630–638, 2011.

[137] C. Taslim, Z. Chen, K. Huang, T. H.-M. Huang, Q. Wang, and S. Lin, "Integrated analysis identifies a class of androgen-responsive genes regulated by short combinatorial long-range mechanism facilitated by ctcf," *Nucleic acids research*, vol. 40, no. 11, pp. 4754–4764, 2012.

[138] X. Xie, T. S. Mikkelsen, A. Gnirke, K. Lindblad-Toh, M. Kellis, and E. S. Lander, "Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of ctcf insulator sites," *Proceedings of the National Academy of Sciences*, vol. 104, no. 17, pp. 7145–7150, 2007.

[139] J. Yang, E. Ramos, and V. G. Corces, "The beaf-32 insulator coordinates genome organization and function during the evolution of drosophila species," *Genome research*, vol. 22, no. 11, pp. 2199–2207, 2012.

[140] M. Merkenschlager and D. T. Odom, "Ctcf and cohesin: linking gene regulatory elements with their targets," *Cell*, vol. 152, no. 6, pp. 1285–1297, 2013.

[141] M. Dluhosova, N. Curik, J. Vargova, A. Jonasova, T. Zikmund, and T. Stopka, "Epigenetic control of spi1 gene by ctcf and iswi atpase smarca5," *PloS one*, vol. 9, no. 2, p. e87448, 2014.

[142] Z. Liu, D. R. Scannell, M. B. Eisen, and R. Tjian, "Control of embryonic stem cell lineage commitment by core promoter factor, taf3," *Cell*, vol. 146, no. 5, pp. 720–731, 2011.

[143] Y. Guo, K. Monahan, H. Wu, J. Gertz, K. E. Varley, W. Li, R. M. Myers, T. Maniatis, and Q. Wu, "Ctcf/cohesin-mediated dna looping is required for protocadherin α promoter choice," *Proceedings of the National Academy of Sciences*, vol. 109, no. 51, pp. 21081–21086, 2012.

[144] T. Hirayama, E. Tarusawa, Y. Yoshimura, N. Galjart, and T. Yagi, "Ctcf is required for neural development and stochastic expression of clustered pcdh genes in neurons," *Cell reports*, vol. 2, no. 2, pp. 345–357, 2012.

[145] K. Monahan, N. D. Rudnick, P. D. Kehayova, F. Pauli, K. M. Newberry, R. M. Myers, and T. Maniatis, "Role of ccctc binding factor (ctcf) and cohesin in the generation of single-cell diversity of protocadherin-α gene expression," *Proceedings of the National Academy of Sciences*, vol. 109, no. 23, pp. 9125–9130, 2012.

[146] P. Majumder, J. A. Gomez, B. P. Chadwick, and J. M. Boss, "The insulator factor ctcf controls mhc class ii gene expression and is required for the formation of long-distance chromatin interactions," *The Journal of experimental medicine*, vol. 205, no. 4, pp. 785–798, 2008.

[147] P. Majumder and J. M. Boss, "Ctcf controls expression and chromatin architecture of the human major histocompatibility complex class ii locus," *Molecular and cellular biology*, vol. 30, no. 17, pp. 4211–4223, 2010.

[148] Y. Shen, F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, *et al.*, "A map of the cis-regulatory sequences in the mouse genome," *Nature*, vol. 488, no. 7409, pp. 116–120, 2012.

[149] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, *et al.*, "Spatial partitioning of the regulatory landscape of the x-inactivation centre," *Nature*, vol. 485, no. 7398, pp. 381–385, 2012.

[150] V. C. Seitan, A. J. Faure, Y. Zhan, R. P. McCord, B. R. Lajoie, E. Ing-Simmons, B. Lenhard, L. Giorgetti, E. Heard, A. G. Fisher, *et al.*, "Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments," *Genome research*, vol. 23, no. 12, pp. 2066–2077, 2013.

[151] S. Sofueva, E. Yaffe, W.-C. Chan, D. Georgopoulou, M. V. Rudan, H. Mira-Bontenbal, S. M. Pollard, G. P. Schroth, A. Tanay, and S. Hadjur, "Cohesin-mediated interactions organize chromosomal domain architecture," *The EMBO journal*, vol. 32, no. 24, pp. 3119–3129, 2013.

[152] J. Zuin, J. R. Dixon, M. I. van der Reijden, Z. Ye, P. Kolovos, R. W. Brouwer, M. P. van de Corput, H. J. van de Werken, T. A. Knoch, W. F. van IJcken, *et al.*, "Cohesin and ctcf differentially affect chromatin architecture and gene expression in human cells," *Proceedings of the National Academy of Sciences*, vol. 111, no. 3, pp. 996–1001, 2014.

[153] P. Collas and J. A. Dahl, "Chop it, chip it, check it: the current status of chromatin immunoprecipitation," *Front Biosci*, vol. 13, no. 17, pp. 929–943, 2008.

[154] L. P. O'Neill and B. M. Turner, "Immunoprecipitation of native chromatin: Nchip," *Methods*, vol. 31, no. 1, pp. 76–82, 2003.

[155] L. P. O'Neill and B. M. Turner, "Immunoprecipitation of chromatin.," *Methods in enzymology*, vol. 274, pp. 189–197, 1995.

[156] P. Collas, "The current state of chromatin immunoprecipitation," *Molecular biotechnology*, vol. 45, no. 1, pp. 87–100, 2010.

[157] P. Collas, *The state-of-the-art of chromatin immunoprecipitation*. Springer, 2009.

[158] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-dna interactions," *Science*, vol. 316, no. 5830, pp. 1497–1502, 2007.

[159] T. S. Furey, "Chip–seq and beyond: new and improved methodologies to detect and characterize protein–dna interactions," *Nature Reviews Genetics*, vol. 13, no. 12, pp. 840–852, 2012.

[160] M. B. Rye, P. Sætrom, and F. Drabløs, "A manually curated chip-seq benchmark demonstrates room for improvement in current peak-finder programs," *Nucleic acids research*, p. gkq1187, 2010.

[161] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, "Capturing chromosome conformation," *Science*, vol. 295, no. 5558, pp. 1306–1311, 2002.

[162] D. Barker, M. Schafer, and R. White, "Restriction sites containing cpg show a higher frequency of polymorphism in human dna," *Cell*, vol. 36, no. 1, pp. 131–138, 1984.

[163] J. Dekker, M. A. Marti-Renom, and L. A. Mirny, "Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data," *Nature Reviews Genetics*, vol. 14, no. 6, pp. 390–403, 2013.

[164] E. de Wit and W. de Laat, "A decade of 3c technologies: insights into nuclear organization," *Genes & development*, vol. 26, no. 1, pp. 11–24, 2012.

[165] J. Dekker, "The three'c's of chromosome conformation capture: controls, controls, controls," *Nature methods*, vol. 3, no. 1, pp. 17–21, 2006.

[166] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, *et al.*, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *science*, vol. 326, no. 5950, pp. 289–293, 2009.

[167] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, *et al.*, "A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, vol. 159, no. 7, pp. 1665–1680, 2014.

[168] T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, and P. Fraser, "Single-cell hi-c reveals cell-to-cell variability in chromosome structure," *Nature*, vol. 502, no. 7469, pp. 59–64, 2013.

[169] M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, *et al.*, "An oestrogen-receptor-&agr;-bound human chromatin interactome," *Nature*, vol. 462, no. 7269, pp. 58–64, 2009.

[170] S. A. Sajan and R. D. Hawkins, "Methods for identifying higher-order chromatin structure," *Annual review of genomics and human genetics*, vol. 13, pp. 59–82, 2012.

[171] Y. Chu and D. R. Corey, "Rna sequencing: platform selection, experimental design, and data interpretation," *Nucleic acid therapeutics*, vol. 22, no. 4, pp. 271–274, 2012.

[172] N. T. Ingolia, G. A. Brar, S. Rouskin, A. M. McGeachy, and J. S. Weissman, "The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mrna fragments," *Nature protocols*, vol. 7, no. 8, pp. 1534–1550, 2012.

[173] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by rna-seq," *Nature methods*, vol. 5, no. 7, pp. 621–628, 2008.

[174] N. Cloonan, A. R. Forrest, G. Kolle, B. B. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, *et al.*, "Stem cell transcriptome profiling via massive-scale mrna sequencing," *Nature methods*, vol. 5, no. 7, pp. 613–619, 2008.

[175] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, "The transcriptional landscape of the yeast genome defined by rna sequencing," *Science*, vol. 320, no. 5881, pp. 1344–1349, 2008.

[176] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14," *cell*, vol. 75, no. 5, pp. 843–854, 1993.

[177] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun, "The 21-nucleotide let-7 rna regulates developmental timing in caenorhabditis elegans," *nature*, vol. 403, no. 6772, pp. 901–906, 2000.

[178] A. E. Pasquinelli, B. J. Reinhart, F. Slack, M. Q. Martindale, M. I. Kuroda, B. Maller, D. C. Hayward, E. E. Ball, B. Degnan, P. Müller, *et al.*, "Conservation of the sequence and temporal expression of let-7 heterochronic regulatory rna," *Nature*, vol. 408, no. 6808, pp. 86–89, 2000.

[179] V. N. Kim, J. Han, and M. C. Siomi, "Biogenesis of small rnas in animals," *Nature reviews Molecular cell biology*, vol. 10, no. 2, pp. 126–139, 2009.

[180] A. Ralston, "Simultaneous gene transcription and translation in bacteria," *Nature Education*, vol. 1, no. 1, p. 4, 2008.

[181] R. D. Morin, M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. J. Pugh, H. McDonald, R. Varhol, S. J. Jones, and M. A. Marra, "Profiling the hela s3 transcriptome using randomly primed cdna and massively parallel short-read sequencing," *Biotechniques*, vol. 45, no. 1, p. 81, 2008.

[182] J. E. Wilusz, H. Sunwoo, and D. L. Spector, "Long noncoding rnas: functional surprises from the rna world," *Genes & development*, vol. 23, no. 13, pp. 1494–1504, 2009.

[183] S. Blackshaw, S. Harpavat, J. Trimarchi, L. Cai, H. Huang, W. P. Kuo, G. Weber, K. Lee, R. E. Fraioli, S.-H. Cho, *et al.*, "Genomic analysis of mouse retinal development," 2004.

[184] M. E. Dinger, P. P. Amaral, T. R. Mercer, K. C. Pang, S. J. Bruce, B. B. Gardiner, M. E. Askarian-Amiri, K. Ru, G. Soldà, C. Simons, *et al.*, "Long noncoding rnas in mouse embryonic stem cell pluripotency and differentiation," *Genome research*, vol. 18, no. 9, pp. 1433–1445, 2008.

[185] T. Ravasi, H. Suzuki, K. C. Pang, S. Katayama, M. Furuno, R. Okunishi, S. Fukuda, K. Ru, M. C. Frith, M. M. Gongora, *et al.*, "Experimental validation of the regulated expression of large numbers of non-coding rnas from the mouse genome," *Genome research*, vol. 16, no. 1, pp. 11–19, 2006.

[186] T. R. Mercer, M. E. Dinger, S. M. Sunkin, M. F. Mehler, and J. S. Mattick, "Specific expression of long noncoding rnas in the mouse brain," *Proceedings of the National Academy of Sciences*, vol. 105, no. 2, pp. 716–721, 2008.

[187] M. Szymanski, M. Z. Barciszewska, V. A. Erdmann, and J. Barciszewski, "A new frontier for molecular medicine: noncoding rnas," *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, vol. 1756, no. 1, pp. 65–75, 2005.

[188] K. V. Prasanth and D. L. Spector, "Eukaryotic regulatory rnas: an answer to the 'genome complexity'conundrum," *Genes & development*, vol. 21, no. 1, pp. 11–42, 2007.

[189] C. A. Maher, C. Kumar-Sinha, X. Cao, S. Kalyana-Sundaram, B. Han, X. Jing, L. Sam, T. Barrette, N. Palanisamy, and A. M. Chinnaiyan, "Transcriptome sequencing to detect gene fusions in cancer," *Nature*, vol. 458, no. 7234, pp. 97–101, 2009.

[190] M. Burrows and D. J. Wheeler, "A block-sorting lossless data compression algorithm," 1994.

[191] P. Ferragina and G. Manzini, "Opportunistic data structures with applications," in *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pp. 390–398, IEEE, 2000.

[192] N. A. Fonseca, J. Rung, A. Brazma, and J. C. Marioni, "Tools for mapping high-throughput sequencing data," *Bioinformatics*, p. bts605, 2012.

[193] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nature methods*, vol. 9, no. 4, pp. 357–359, 2012.

[194] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows–wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[195] F. M. Pauler, M. A. Sloane, R. Huang, K. Regha, M. V. Koerner, I. Tamir, A. Sommer, A. Aszodi, T. Jenuwein, and D. P. Barlow, "H3k27me3 forms blocs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome," *Genome research*, vol. 19, no. 2, pp. 221–233, 2009.

[196] M. D. Young, T. A. Willson, M. J. Wakefield, E. Trounson, D. J. Hilton, M. E. Blewitt, A. Oshlack, and I. J. Majewski, "Chip-seq analysis reveals distinct h3k27me3 profiles that correlate with transcriptional activity," *Nucleic acids research*, vol. 39, no. 17, pp. 7415–7427, 2011.

[197] J. P. Reddington, S. M. Perricone, C. E. Nestor, J. Reichmann, N. A. Youngson, M. Suzuki, D. Reinhardt, D. S. Dunican, J. G. Prendergast, H. Mjoseng, *et al.*, "Redistribution of h3k27me3 upon dna hypomethylation results in de-repression of polycomb target genes," *Genome Biol*, vol. 14, no. 3, p. R25, 2013.

[198] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, *et al.*, "Model-based analysis of chip-seq (macs)," *Genome biology*, vol. 9, no. 9, p. R137, 2008.

[199] C. Zang, D. E. Schones, C. Zeng, K. Cui, K. Zhao, and W. Peng, "A clustering approach for identification of enriched domains from histone modification chip-seq data," *Bioinformatics*, vol. 25, no. 15, pp. 1952–1958, 2009.

[200] J. Wang, V. V. Lunyak, and I. K. Jordan, "Broadpeak: a novel algorithm for identifying broad peaks in diffuse chip-seq datasets," *Bioinformatics*, p. bts722, 2013.

[201] Q. Song and A. D. Smith, "Identifying dispersed epigenomic domains from chip-seq data," *Bioinformatics*, vol. 27, no. 6, pp. 870–871, 2011.

[202] A. Agresti and B. A. Coull, "Approximate is better than "exact" for interval estimation of binomial proportions," *The American Statistician*, vol. 52, no. 2, pp. 119–126, 1998.

[203] W. L. Ruzzo and M. Tompa, "A linear time algorithm for finding all maximal scoring subsequences.," in *ISMB*, vol. 99, pp. 234–241, 1999.

[204] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[205] B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, *et al.*, "The nih roadmap epigenomics mapping consortium," *Nature biotechnology*, vol. 28, no. 10, pp. 1045–1048, 2010.

[206] M. Lizio, J. Harshbarger, H. Shimoji, J. Severin, T. Kasukawa, S. Sahin, I. Abugessaisa, S. Fukuda, F. Hori, S. Ishikawa-Kato, *et al.*, "Gateways to the fantom5 promoter level mammalian expression atlas," *Genome biology*, vol. 16, no. 1, p. 22, 2015.

[207] B. D. Strahl and C. D. Allis, "The language of covalent histone modifications," *Nature*, vol. 403, no. 6765, pp. 41–45, 2000.

[208] G. Hon, B. Ren, and W. Wang, "Chromasig: a probabilistic approach to finding common chromatin signatures in the human genome," *PLoS Comput Biol*, vol. 4, no. 10, p. e1000201, 2008.

[209] J. Ernst and M. Kellis, "Discovery and characterization of chromatin states for systematic annotation of the human genome," *Nature biotechnology*, vol. 28, no. 8, pp. 817–825, 2010.

[210] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny, "Iterative correction of hi-c data reveals hallmarks of chromosome organization," *Nature methods*, vol. 9, no. 10, pp. 999–1003, 2012.

[211] Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble, "A three-dimensional model of the yeast genome," *Nature*, vol. 465, no. 7296, pp. 363–367, 2010.

[212] F. Ay, T. L. Bailey, and W. S. Noble, "Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts," *Genome research*, vol. 24, no. 6, pp. 999–1011, 2014.

[213] E. Yaffe and A. Tanay, "Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture," *Nature genetics*, vol. 43, no. 11, pp. 1059–1065, 2011.

[214] J. Paulsen, E. A. Rødland, L. Holden, M. Holden, and E. Hovig, "A statistical model of chia-pet data for accurate detection of chromatin 3d interactions," *Nucleic acids research*, vol. 42, no. 18, pp. e143–e143, 2014.

[215] R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, *et al.*, "An atlas of active enhancers across human cell types and tissues," *Nature*, vol. 507, no. 7493, pp. 455–461, 2014.

[216] B. He, C. Chen, L. Teng, and K. Tan, "Global view of enhancer–promoter interactome in human cells," *Proceedings of the National Academy of Sciences*, vol. 111, no. 21, pp. E2191–E2199, 2014.

[217] H. Kikuta, M. Laplante, P. Navratilova, A. Z. Komisarczuk, P. G. Engström, D. Fredman, A. Akalin, M. Caccamo, I. Sealy, K. Howe, *et al.*, "Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates," *Genome research*, vol. 17, no. 5, pp. 545–555, 2007.

[218] D. M. Larkin, G. Pape, R. Donthu, L. Auvil, M. Welge, and H. A. Lewin, "Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories," *Genome research*, vol. 19, no. 5, pp. 770–777, 2009.

[219] R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, *et al.*, "The accessible chromatin landscape of the human genome," *Nature*, vol. 489, no. 7414, pp. 75–82, 2012.

[220] O. Corradin, A. Saiakhova, B. Akhtar-Zaidi, L. Myeroff, J. Willis, R. Cowper-Sal, M. Lupien, S. Markowitz, P. C. Scacheri, *et al.*, "Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits," *Genome research*, vol. 24, no. 1, pp. 1–13, 2014.

[221] J. Schug, W.-P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan, and C. J. Stoeckert, "Promoter features related to tissue specificity as measured by shannon entropy," *Genome biology*, vol. 6, no. 4, p. R33, 2005.

[222] S. Roy, A. F. Siahpirani, D. Chasman, S. Knaack, F. Ay, R. Stewart, M. Wilson, and R. Sridharan, "A predictive modeling approach for cell line-specific long-range regulatory interactions," *Nucleic acids research*, p. gkv865, 2015.

[223] M. Meyerson, S. Gabriel, and G. Getz, "Advances in understanding cancer genomes through second-generation sequencing," *Nature Reviews Genetics*, vol. 11, no. 10, pp. 685–696, 2010.

[224] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome Biol*, vol. 14, no. 4, p. R36, 2013.

[225] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. Van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature biotechnology*, vol. 28, no. 5, pp. 511–515, 2010.

[226] R. P. Dilworth, "A decomposition theorem for partially ordered sets," *Annals of Mathematics*, pp. 161–166, 1950.

[227] V. M. Kvam, P. Liu, and Y. Si, "A comparison of statistical methods for detecting differentially expressed genes from rna-seq data," *American journal of botany*, vol. 99, no. 2, pp. 248–256, 2012.

[228] I. Nookaew, M. Papini, N. Pornputtpong, G. Scalcinati, L. Fagerberg, M. Uhlén, and J. Nielsen, "A comprehensive comparison of rna-seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in saccharomyces cerevisiae," *Nucleic acids research*, p. gks804, 2012.

[229] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[230] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome biol*, vol. 11, no. 10, p. R106, 2010.

[231] J. Li and R. Tibshirani, "Finding consistent patterns: a nonparametric approach for identifying differential expression in rna-seq data," *Statistical methods in medical research*, vol. 22, no. 5, pp. 519–536, 2013.

[232] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter, "Differential analysis of gene regulation at transcript resolution with rna-seq," *Nature biotechnology*, vol. 31, no. 1, pp. 46–53, 2013.

[233] A. Oshlack, M. D. Robinson, M. D. Young, *et al.*, "From rna-seq reads to differential expression results," *Genome biol*, vol. 11, no. 12, p. 220, 2010.

[234] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

[235] F. Seyednasrollah, A. Laiho, and L. L. Elo, "Comparison of software packages for detecting differential expression in rna-seq studies," *Briefings in bioinformatics*, vol. 16, no. 1, pp. 59–70, 2015.

[236] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, "Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks," *Nature protocols*, vol. 7, no. 3, pp. 562–578, 2012.

[237] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, *et al.*, "Bioconductor: open software development for computational biology and bioinformatics," *Genome biology*, vol. 5, no. 10, p. R80, 2004.

[238] R. C. Team, "R: A language and environment for statistical computing. vienna, austria; 2014," *URL http://www. R-project. org*, 2015.

[239] R. Ihaka and R. Gentleman, "R: a language for data analysis and graphics," *Journal of computational and graphical statistics*, vol. 5, no. 3, pp. 299–314, 1996.

[240] H. Wickham, *ggplot2: elegant graphics for data analysis*. Springer Science & Business Media, 2009.

[241] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov, "Integrative genomics viewer," *Nature biotechnology*, vol. 29, no. 1, pp. 24–26, 2011.

[242] A. C. Boquest, A. Shahdadfar, K. Frønsdal, O. Sigurjonsson, S. H. Tunheim, P. Collas, and J. E. Brinchmann, "Isolation and transcription profiling of purified uncultured human stromal stem cells: alteration of gene expression after in vitro cell culture," *Molecular biology of the cell*, vol. 16, no. 3, pp. 1131–1141, 2005.

[243] J.-L. Boulland, M. Mastrangelopoulou, A. C. Boquest, R. Jakobsen, A. Noer, J. C. Glover, and P. Collas, "Epigenetic regulation of nestin expression during neurogenic differentiation of adipose tissue stem cells," *Stem cells and development*, vol. 22, no. 7, pp. 1042–1052, 2012.

[244] A. Mammana and H.-R. Chung, "Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome," *Genome biology*, vol. 16, no. 1, pp. 1–12, 2015.

[245] M. V. Rudan, C. Barrington, S. Henderson, C. Ernst, D. T. Odom, A. Tanay, and S. Hadjur, "Comparative hi-c reveals that ctcf underlies evolution of chromosomal domain architecture," *Cell reports*, vol. 10, no. 8, pp. 1297–1309, 2015.

[246] Y. Guo, Q. Xu, D. Canzio, J. Shou, J. Li, D. U. Gorkin, I. Jung, H. Wu, Y. Zhai, Y. Tang, *et al.*, "Crispr inversion of ctcf sites alters genome topology and enhancer/promoter function," *Cell*, vol. 162, no. 4, pp. 900–910, 2015.

[247] N. Ahituv, S. Prabhakar, F. Poulin, E. M. Rubin, and O. Couronne, "Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny," *Human molecular genetics*, vol. 14, no. 20, pp. 3057–3063, 2005.

[248] B. D. Pope, T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, D. L. Vera, Y. Wang, R. S. Hansen, T. K. Canfield, *et al.*, "Topologically associating domains are stable units of replication-timing regulation," *Nature*, vol. 515, no. 7527, pp. 402–405, 2014.

[249] Y. C. Lin, C. Benner, R. Mansson, S. Heinz, K. Miyazaki, M. Miyazaki, V. Chandra, C. Bossen, C. K. Glass, and C. Murre, "Global changes in the nuclear positioning of genes and intra-and interdomain genomic interactions that orchestrate b cell fate," *Nature immunology*, vol. 13, no. 12, pp. 1196–1204, 2012.

[250] J. Su, S. A. Teichmann, and T. A. Down, "Assessing computational methods of cis-regulatory module prediction," *PLoS Comput Biol*, vol. 6, no. 12, p. e1001020, 2010.

[251] X.-Y. Li, S. MacArthur, R. Bourgon, D. Nix, D. A. Pollard, V. N. Iyer, A. Hechmer, L. Simirenko, M. Stapleton, C. L. Luengo Hendriks, *et al.*, "Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm," *PLoS Biol*, vol. 6, no. 2, p. e27, 2008.

[252] K. Y. Yip, C. Cheng, and M. Gerstein, "Machine learning and genome annotation: a match meant to be," *Genome Biol*, vol. 14, no. 5, p. 205, 2013.

[253] M. J. Blow, D. J. McCulley, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, *et al.*, "Chip-seq identification of weakly conserved heart enhancers," *Nature genetics*, vol. 42, no. 9, pp. 806–810, 2010.

[254] D. May, M. J. Blow, T. Kaplan, D. J. McCulley, B. C. Jensen, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, *et al.*, "Large-scale discovery of enhancers from human heart tissue," *Nature genetics*, vol. 44, no. 1, pp. 89–93, 2012.

[255] N. Rajagopal, W. Xie, Y. Li, U. Wagner, W. Wang, J. Stamatoyannopoulos, J. Ernst, M. Kellis, and B. Ren, "Rfecs: a random-forest based algorithm for enhancer identification from chromatin state," *PLoS Comput. Biol*, vol. 9, no. 3, p. e1002968, 2013.

[256] T.-K. Kim, M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, *et al.*, "Widespread transcription at neuronal activity-regulated enhancers," *Nature*, vol. 465, no. 7295, pp. 182–187, 2010.

[257] F. Lai, A. Gardini, A. Zhang, and R. Shiekhattar, "Integrator mediates the biogenesis of enhancer rnas," *Nature*, vol. 525, no. 7569, pp. 399–403, 2015.

[258] D. Lara-Astiaso, A. Weiner, E. Lorenzo-Vivas, I. Zaretsky, D. A. Jaitin, E. David, H. Keren-Shaul, A. Mildner, D. Winter, S. Jung, *et al.*, "Chromatin state dynamics during blood formation," *Science*, vol. 345, no. 6199, pp. 943–949, 2014.

[259] N. Nègre, C. D. Brown, L. Ma, C. A. Bristow, S. W. Miller, U. Wagner, P. Kheradpour, M. L. Eaton, P. Loriaux, R. Sealfon, *et al.*, "A cis-regulatory map of the drosophila genome," *Nature*, vol. 471, no. 7339, pp. 527–531, 2011.

[260] A. Noer, L. C. Lindeman, and P. Collas, "Histone h3 modifications associated with differentiation and long-term culture of mesenchymal adipose stem cells," *Stem cells and development*, vol. 18, no. 5, pp. 725–736, 2009.

[261] L. A. Meza-Zepeda, A. Noer, J. A. Dahl, F. Micci, O. Myklebost, and P. Collas, "High-resolution analysis of genetic stability of human adipose tissue stem cells cultured to senescence," *Journal of cellular and molecular medicine*, vol. 12, no. 2, pp. 553–563, 2008.

[262] Y. Sato, M. Mukai, J. Ueda, M. Muraki, T. J. Stasevich, N. Horikoshi, T. Kujirai, H. Kita, T. Kimura, S. Hira, *et al.*, "Genetically encoded system to track histone modification in vivo," *Scientific reports*, vol. 3, 2013.

[263] T. Muramoto, D. Cannon, M. Gierliński, A. Corrigan, G. J. Barton, and J. R. Chubb, "Live imaging of nascent rna dynamics reveals distinct types of transcriptional pulse regulation," *Proceedings of the National Academy of Sciences*, vol. 109, no. 19, pp. 7350–7355, 2012.

[264] C. Y. Jao and A. Salic, "Exploring rna transcription and turnover in vivo by using click chemistry," *Proceedings of the National Academy of Sciences*, vol. 105, no. 41, pp. 15779–15784, 2008.

[265] B. Chen, L. A. Gilbert, B. A. Cimini, J. Schnitzbauer, W. Zhang, G.-W. Li, J. Park, E. H. Blackburn, J. S. Weissman, L. S. Qi, *et al.*, "Dynamic imaging of genomic loci in living human cells by an optimized crispr/cas system," *Cell*, vol. 155, no. 7, pp. 1479–1491, 2013.